计算机科学
COMPUTER SCIENCE

基于特征再抽象(FRA)的多元时序预测方法

王昊, 周建涛, 郝昕毓, 王飞宇

引用本文

王昊, 周建涛, 郝昕毓, 王飞宇. 基于特征再抽象(FRA)的多元时序预测方法[J]. 计算机科学, 2023, 50(11A): 221100144-8.

WANG Hao, ZHOU Jiantao, HAO Xinyu, WANG Feiyu. Multivariate Time Series Forecasting Method Based on FRA [J]. Computer Science, 2023, 50(11A): 221100144-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100144

基于特征再抽象(FRA)的多元时序预测方法

王 昊 周建涛 郝昕毓 王飞宇
内蒙古大学计算机学院 呼和浩特 010021
蒙古文智能信息处理技术国家地方联合工程研究中心 呼和浩特 010021
生态大数据教育部工程研究中心 呼和浩特 010021
内蒙古自治区云计算与服务软件工程实验室 呼和浩特 010021
内蒙古自治区社会计算与数据处理重点实验室 呼和浩特 010021
内蒙古自治区大数据分析技术工程实验室 呼和浩特 010021
内蒙古自治区纪检监察大数据重点实验室 呼和浩特 010021
内蒙古自治区大数据分析技术工程实验室 呼和浩特 010021
(2892733460@qq.com)

摘 要 科技领域的衍生行业因普遍存在强时间约束的特性而累积了海量的高维时间序列数据,严峻的数据压力导致传统的数据建模预测方法受制于数据规模和属性维度。支撑高质量的服务对大数据智能预测技术提出了更高的要求,如何在数据层面上实现预测性能的提升是现阶段亟待解决的主要问题。针对上述问题,提出了针对多元时序数据的特征再抽象(Feature Re-Abstraction,FRA)算法,首先通过 RobustSTL 分解算法提取趋势性和季节性特征(Trend and Seasonality Features,TSFs),实现多元数据的特征二阶抽象,以"抽象即特征"替代传统"标签即特征"的提取策略,再通过 Pearson 相关系数的运算结果评估再抽象技术捕捉的 TSFs 与目标参数间的相关强度,证实 TSF 的数据价值。在 FRA 算法的基础上结合深度学习模型构建基于数据驱动的多元时序预测算法,通过预测效果验证 FRA 算法的有效性。实验结果表明,引入 TSFs 作为数据驱动模型的训练向量能够兼具数据降维、降噪及强相关特性地维持,从而避免模型过拟合并缓解模型欠拟合,提高时序预测算法的准确性和鲁棒性。

关键词:多元时序数据;多元时序预测算法;特征再抽象;趋势性和季节性特征;相关性评估
中图法分类号 TP311.1

Multivariate Time Series Forecasting Method Based on FRA

WANG Hao,ZHOU Jiantao,HAO Xinyu and WANG Feiyu
College of Computer Science,Inner Mongolia University,Hohhot 010021,China
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian,Hohhot 010021,China
Engineering Research Center of Ecological Big Data,Ministry of Education,Hohhot 010021,China
Inner Mongolia Engineering Laboratory for Cloud Computing and Service Software,Hohhot 010021,China
Inner Mongolia Key Laboratory of Social Computing and Data Processing,Hohhot 010021,China
Inner Mongolia Engineering Laboratory for Big Data Analysis Technology,Hohhot 010021,China
Inner Mongolia Key Laboratory of Discipline Inspection and Supervision Big Data,Hohhot 010021,China
Inner Mongolia Big Data Analysis Technology Engineering Laboratory,Hohhot 010021,China

Abstract Derivative industries in the field of science and technology have accumulated a large amount of high-dimensional time series data due to the general existence of strong time constraints. Severe data pressure makes traditional data modeling and prediction methods limited by data scale and attribute dimensions. Services supporting high-quality put forward higher requirements for big data intelligent prediction technology. How to improve the prediction performance at the data level is a main problem that needs to be solved urgently at this stage. Combined with the above problems,a feature re-abstraction(FRA) algorithm for multivariate time series data is proposed. First,the RobustSTL decomposition algorithm is used to extract trend and seasonality features(TSFs),realize the second-order abstraction of features of multivariate data,and replace the traditional extraction strategy of

基金项目:国家自然科学基金(62162046);内蒙古科技攻关项目(2021GG0155);内蒙古自然科学基金重大项目(2019ZD15);内蒙古自然科学基金(2019GG372);内蒙古大学/内蒙古自治区研究生科研创新项目(11200-121024)
This work was supported by the National Natural Science Foundation of China(62162046),Inner Mongolia Science and Technology Research Project(2021GG0155),Major Programs of Inner Mongolia Natural Science Foundation(2019ZD15),Inner Mongolia Natural Science Foundation(2019GG372) and Inner Mongolia University/Inner Mongolia Autonomous Region Graduate Scientific Research Innovation Project(11200-121024).
通信作者:周建涛(cszjtao@imu.edu.cn)

"labels are features" with "abstract is features". Then, the correlation strength between the TSFs captured by the re-abstract technology and the target parameters is evaluated by the calculation result of the Pearson correlation coefficient, which confirms the data value of the TSF. On the basis of FRA algorithm, combined with deep learning model, a data-driven multivariate time series prediction algorithm is constructed, and the effectiveness of FRA algorithm is verified by the prediction effect. Experimental results show that the introduction of TSFs as the training vector of the data-driven model can maintain the characteristics of data dimensionality reduction, noise reduction and strong correlation, so as to avoid model overfitting and alleviate model underfitting, and improve the accuracy and robustness of time series prediction algorithms.

Keywords Multivariate time series data, Multivariate time series forecasting algorithms, Feature re-abstraction(FRA), Trend and seasonality feature(TSF), Correlation assessment

大数据时代，工程和科学领域的信息化数据呈现几何级增长，给智能分析任务造成了严峻的数据压力。现实场景下，金融、工业、医学乃至天文等科技领域都存在着海量的高维时序数据，衍生出了广泛的行业应用需求，迫使研究人员需要对多元序列信息进行深入挖掘与分析来支撑高质量的智能服务，如金融行业分析、设备异常检测、基因序列探索及静态行为识别等[1]。针对多元时序数据的研究工作需利用大数据挖掘技术，以确保捕获到具有时间戳特性的潜在价值信息，并基于分类、聚类及预测等数据应用技术呈现信息价值，促进技术在科研领域的创新。

其中，时序预测能够依据时间依赖特性学习历史属性与未来数值间的映射关系，提升预测效果的同时赋予结果时效性，对于辅助决策、优化资源分配、提前采取止损措施等具有基础且重要的意义[1]。针对多元时序预测问题，众多学者基于不同研究视角贡献了许多可行性方法，绝大多数都是通过构建有效的多元时序预测模型以改进预测性能，分为传统的线性统计类方法[2-4]和机器学习方法[5-14]。以自回归模型(Autoregressive, AR)和自回归滑动平均模型(Autoregressive Moving Average, ARMA)为代表的统计类模型主要用于解决平稳的多元时序数据，但现实应用场景下，趋势性、波动性和周期性等数据特征普遍存在大幅度变化，导致回归模型的预测效果急剧下滑。结合传统方法存在的问题，基于机器学习的时序预测算法被提出并成为目前主流的研究内容，在实际应用中表现出更优秀的预测性能。

大数据模式下，基于机器学习的多元时序预测方法需要遵循数据驱动的执行策略，学习观测序列的辨识性特征，建模特征和未来时间戳对应数值间的映射关系[4]。对于基于机器学习的多元时序预测研究，学者们主要在特征提取和数据建模两方面探索科研突破，从而优化和改善多元时序预测算法。随着科研进度的推进，以神经网络为核心的深度学习算法[5-7,9-12,14]推动着数据建模的研究工作达到了较为成熟的阶段，尤其在应对规模庞大和复杂的工况时取得了更为瞩目的成就。尽管基于机器学习的多元时序预测方法在数据建模角度的研究已较为成熟，可以通过优化模型逐层构建和处理的机制实现预测效果的提升，但特征提取角度的创新和改进才是能够从根源上突破性能瓶颈的关键方法。

针对特征提取的研究工作，多元时序数据的属性维度和高频噪音是在数据层面上限制准确率和鲁棒性的直接因素。为了解决大规模多元时序数据所面临的提取难题，本文针对基于机器学习的时序预测的整体性能，提出了特征再抽象(FRA)技术对观测数据执行抽象操作，得到能够拟合多元序列演变特性的序列变化特征。FRA 方法严格遵循先抽象后评估的渐进式执行机制，首先基于 RobustSTL 时间分解算法对多元时序数据执行多变量特征提取，捕获能够反映时间戳特性的季节性和趋势性两种序列表示形式，兼具序列降维和特征强相关特性，并通过分解数据残差的操作剔除多元时序数据的高频噪音。抽象步骤主要承担特征在属性维度和噪音干扰方面的优化工作，而评估步骤是利用皮尔逊相关性系数来衡量特征向量与目标参数的相关强度，依据参数的运算结果评定特征能否作为深度学习模型进行无监督学习的输入量。FRA 方法作用下，高维度、大规模时序数据在实现属性降维和数据降噪的基础上保证了相关强度，在数据层面提升了多元时序预测的准确性和鲁棒性。

基于 FRA 捕获季节性和趋势性特征(TSFs)，结合以 LSTM 网络为代表的多元时序深度模型构建数据驱动的算法框架，针对具备时间戳特性的锂电池演化数据提出基于 FRA 的多元时序预测算法。其研究思路是基于 7 组真实的锂电池数据集设计实验，通过提取表征数据趋势性、周期性及波动性的时序变化特征构建高质量的特征向量，借助 LSTM 网络在预测性能上的均衡表现，实现锂电池的全周期属性值预测，其预测数值能够反映出锂电池的损耗情况，可以作为参考指标支撑产品质量检测、报废周期评估和异常检测及安全预警等服务[5]。

本文第 1 章主要针对目前研究领域内的相关工作进行介绍；第 2 章主要针对多元时序预测提出的方法进行介绍；第 3 章是实验设计和效果展示；最后总结全文。

1 相关工作

基于机器学习的多元时序预测算法通过严格执行由特征提取到数据建模的数据驱动算法框架，来反映未来时间戳对应的属性数值，能够在适应复杂工况时提供最理想的预测效果。因此，本节将针对机器学习方法的数据建模和特征提取两项技术，进行研究思路和性能表现方面的介绍。

现阶段的核心研究集中于数据建模，即如何构建兼具信息继承和防止过载的深度学习算法，如何利用深度学习算法的自学习能力和数据依赖性推演出序列演化规律。

针对数据建模亟待解决的研究问题而推进的相关研究中，递归神经网络(RNN)是具有里程碑意义的深度学习模型，其权值共享的特性意味着能够提取线性序列中随时间变化的特征，结合模型的记忆性使其具备了序列信息继承的能力，但 RNN 网络的递归原理类似于连续的矩阵乘法，其误差梯度会随着时间步的持续反向传播而容易导致梯度爆炸和

梯度消失的极端化非线性行为，进而影响到预测效果。针对RNN的优化算法中，Shi等[7]提出的长短期记忆网络(Long Short Term Memory network，LSTM)通过赋予门控单元控制内部信息积累和遗忘的能力实现更加灵活的信息继承，既能掌握长距离依赖又能选择性地遗忘信息以防止过载，是目前解决工程性序列型数据任务的主流算法。区别于RNN网络的优化算法，Vaswani[8]提出的Transformer模型，在数据建模阶段大胆撇弃了传统RNN网络基于“记忆”的设计原理，采用能够捕获数据内部相关特性的自注意机制(Self-Attention)构建深度学习模型，在处理并行化数据问题上效果拔群。针对数据建模设计的算法具备数据依赖性，序列的价值密度和去噪程度能够直接影响算法整体性能，如何利用特征提取技术改进数据质量成为了研究热点。

多元时序预测算法的特征提取主要承担数据准备的工作，其研究核心是基于序列关键片段提取具备时间依赖性的可辨识特征向量，如何保证提取分量的价值密度、相关强度及可解释性是目前亟待解决的关键问题。

针对特征提取的研究问题，本文提出了一系列基于序列片段的提取机制，通过挖掘高价值密度区域的片段信息以保证相关强度和可解释性。众多研究中，Ye等提出的Shapelets概念[4]最具代表性，其算法原理是将局部序列视作事件执行相应的操作，充分考虑到序列中局部形状与序列标签的映射关系，为辨识性特征提供可解释性。采用Shapelets算法提取特征的机制是通过枚举策略和信息增益搜索并匹配能够拟合任意局部序列的短模式，间接地捕获到原始序列中高价值密度的特征向量，该方法能够通过调整序列片段而避免噪声干扰掩盖掉序列中高价值的可辨识小序列，但线性的暴力搜索机制使其难以处理多元时序数据。相较于搜索算法，深度学习算法具备更强的局部检索能力。Elhassan等[9]基于传统图像处理和深度学习算法提出了混合特征提取方法，该方法利用CNN的特征融合法提取基于关键区域的深度学习特征。Liu等[10]提出了1D-CNN联合特征提取方法，该方法针对原始序列进行分区裁剪以构建基于分区片段的特征学习空间，并借助1D-CNN网络提取到各区间的代表性特征域。

基于区域片段的相关研究普遍需要执行搜索机制才能够高效地捕获到符合预设定参数的特征向量，基于深度学习算法的特征提取甚至会因为网络架构复杂和分析模块冗余而导致特征提取不彻底、整体任务执行缓慢等问题[11]。

针对上述问题，研究者们逐渐将多元时序预测的研究对象调整为全周期序列数据，并通过属性降维和数据降噪等处理手段保证特征分量的价值密度及相关强度。Zou[12]提出了基于图拉普拉斯变化的特征提取方法，该方法首先依据图拉普拉斯变换理论，针对多元时序数据执行半监督的特征提取，再利用散布矩阵实现监督特征和半监督特征的融合，最后基于图拉普拉斯的相关理论筛选出特征子集。基于特征矩阵再筛选子集的序列降维措施是目前核心的研究方法，它可以在实现缩减规模的同时全面覆盖属性，在数据层面上减少模型负荷。Guo等[13]提出一种基于混合频谱信号编码的低通过滤网格纹理平滑算法，该方法的设计原理是通过特征识别和信息重构的方法剔除高频噪音及弱相关性特征，实验数据证实降低噪音干扰和低价值参数的数据占比可以实现强相关特征分量的信息增强。

针对全周期序列数据的技术研究中，出现了基于序列变化特征的研究分支，研究者们利用趋势、波动和周期等信息来模拟序列的发展规律，依据规律性探索未来时刻的数据变化。Jia等[14]提出的基于多尺度特征提取方法主要用于扩充特征池，通过多尺度卷积模块对全周期序列数据实现时域、频域和时频域特征的无监督提取，其实验效果证实以时域信息为代表的序列变化特征在面向空间信息不足的预测问题时被作为特征向量的可解释性，具备应对高频噪声和失真问题的稳定性。Liu等[15]提出了基于转折点和趋势段的多元时序趋势特征提取算法，主要通过调整序列分段的评定标准提升辨识性趋势特征的精度，该方法首先利用序列拐点定义趋势段，再通过选定极值趋势段来确定全周期序列数据最合理的分段点，将如何评判转折点转向如何评判趋势段，最大程度上避免了陷入局部最优和特征弱相关性的问题，其实验对照结果证实具有趋势演变性质的辨识性特征能够配合算法达到更理想的拟合精度，具备更好的抗噪能力。

总结上述方法可知，基于数据建模的研究成果基本能够保障高性能的预测效果。其中，以LSTM网络为代表的RNN门控算法兼具准确性和稳定性，在多元时序预测的科技领域中应用也最为广泛。基于特征提取的研究内容主要集中于如何保证特征具备高价值密度和强相关特征，但现阶段的研究问题是不能充分考虑到属性维度、数据噪音和序列片段化，导致难以提取到最优特征分量。正如本文引言部分所述，多元时序预测的研究工作呈现多边化发展，其性能提升的关键是基于高质量的数据基础，并结合高性能的深度学习算法。本文提出的基于FRA的多元时序预测算法，首先基于趋势性和季节性特征构建能够模拟序列变化规律的特征向量组，通过充分考虑到序列数据的降维、降噪和相关性保障特征质量，再结合高性能的LSTM网络进行无监督学习，实现兼具准确性和鲁棒性的多元时序预测。

2　基于特征再抽象的多元时序预测算法

相关文献[16-17]评价趋势性和季节性特征(Trend and Seasonality Feature，TSF)是针对时间序列移动方向的一种高层次表现形式，能够表现序列稳定起伏和波动等趋势依赖关系[14]，并依据序列的发展规律探索未来时刻的数据变化。本章着重介绍针对全周期的多元时间序列数据提取低维、低噪且高精度TSFs的特征再抽象(FRA)算法，最后介绍如何基于数据驱动的运行策略联立FRA算法及长短期记忆网络。

2.1　特征再抽象(FRA)算法

本文的设计理念和效果实现是基于先抽象后评估的递进式执行机制，首先基于最小绝对偏差与非局部季节滤波改进RobustSTL[18]算法，用于面向多元高维序列样本执行特征二阶抽象，提取能够映射样本序列涨幅程度及波动频率的低维度特征分量，其次基于残差收敛性和Pearson关联评估设置递归迭代刷新特征分解纯度，以此提出了特征再抽象(FRA)算法。

其中，RobustSTL时间分解算法执行序列型特征的二阶抽象操作，以原始序列的多维标签属性为基础提取能够模拟数据变化规律的序列变换特征，譬如能够表征序列演变涨幅

和变化频率的趋势性和季节性特征(TSF),依据 TSF 构建出式(1)模式下的特征二元组 $F_i=(T_i,S_i)$,其中 T_i 表示趋势性特征,S_i 表示季节性特征,从而实现了特征向量降维;由皮尔逊相关系数(Pearson correlation coefficient)的运算结果评估特征二元组 F 和预期目标参数之间的相关强度,借助相关性判断 F 是否存在因数据维度低而影响到预测效果的可能性,从而决断出导入深度学习算法的特征向量组。

$$F=(T_i,S_i)\mid 1\leqslant i\leqslant n,n\geqslant 0,i,n\in N^* \tag{1}$$

依据特征划分机制,TSF 被划分为无量纲时域特征,此类特征是指以时间属性为变量,用于描绘数据演变波形的向量集,对采集设备的负载和工况变化更具抵抗性,能够通过降低采集环境和关联领域的影响,稳定地表现数据演变波形。依据时间序列特征提取的思路划分,本文提出的二阶抽象特征提取机制符合序列变换特征的提取方法,旨在采用特定算法将原始时间序列转化为另一种表现形式的特征序列,能够有效地应对时间序列中存在的噪声、失真等问题。

针对待处理的原始时间序列,定义 y_t 为 t 时刻的观测数据,具备强依赖于时间属性的特点。定义 T_t 为观测数据的趋势信号(Trend signal),用于描述数据连续增加或减少、曲线上升或下降的涨幅特性;S_t 为季节信号(Seasonal signal),连续的 S_t 能够描述在基线附近波动的周期性模式,通过学习由 T_t 和 S_t 构建的向量组即可获取到原始序列演变模式等关键信息。R_t 表示时间序列数据分解后的残差(Residual),如式(2)所示:

$$R_t=a_t+n_t \tag{2}$$

公式定义残差是由通俗意义上的 n_t 白噪音和 a_t 尖峰或低谷(spike or dip)构成,其因和目标参数具有弱关联特性,故被视为不具备模拟序列演变特性的高频噪音。时间序列 y_t 与趋势项、季节项和残差的加法模型被定义为式(3):

$$y_t=T_t+S_t+R_t,t=1,2,\cdots,N \tag{3}$$

针对基于加法模型的相关研究中,时间序列分解算法(Seasonal and Trend decomposition using Loess, STL)[19] 较传统分解算法对异常值更具鲁棒性,其通过设置内循环(Inner Loop)与外循环(Outer Loop)分别承担趋势拟合与周期分量的计算以及鲁棒局部加权回归权重(Robustness Weight)的调节任务来保证算法具备足够的鲁棒性。STL 算法在分解成分步骤中更具灵活性,季节性成分随时间变化的速度及周期性跨度、趋势性成分的周期平滑度等都可以由用户控制干预,更能适应不同工况条件下采集的时间序列数据。

本文优化的 RobustSTL 时间序列分解算法在保持 STL 算法鲁棒性和灵活性的同时,在成分分解纯度上进行了性能优化。对 TSF 特征而言,特征分解的纯度直接影响到特征二元组和目标参数之间的相关性强度;对高频噪音而言,提升分解纯度意味着基于 RobustSTL 算法的数据预测任务对噪音干扰具备更强的抗性,即使在应对波动性大且突变率高的时间序列数据时也能保证算法鲁棒性。该算法针对 TSF 分量的提取原理可划分为两个阶段:首先利用具有稀疏正则化的最小绝对偏差损失解决多元线性回归问题,通过赋予算法处理趋势和残差突变特性的能力,稳健地提取多元时间序列的趋势性分量;再基于本轮递归循环内提取的趋势性分量应用非局部季节性滤波来提取季节性分量,通过赋予算法处理季节波动和偏移等情况的能力来提升待处理季节性特征的灵活性;通过递归的迭代机制重复刷新趋势性和季节性分量的提取纯度,依据递归的收敛参数判定是否输出高精度的 TSF。

算法 1 详细描述了优化 RobustSTL 时间分解算法的实现机制,算法执行步骤包括双边滤波降噪、LAD 损失回归提取趋势、非局部季节性过滤提取季节以及特征调整。

算法 1 RobustSTL(Sample)

Input: Sample

Output: [Trend, Seasonality, Remainder]

Variables: $\text{Season}_{\text{length}}$

Reg_1 (first-order regularization parameter)

Reg_2 (second-order regularization parameter)

K

H

Hbn_1, Hbn_2

Hbs_1, Hbs_2

1. Initialize parameters: $\text{Reg}_1\leftarrow 10.0$, $\text{Reg}_2\leftarrow 0.5$, $K\leftarrow 2$, $H\leftarrow 5$, $\text{Hbs}_1\leftarrow 50$, Hbs_2 & Hbn_1 & $\text{Hbn}_2\leftarrow 1$
2. while true do
3. Trial$\leftarrow$1
4. $\text{denoise}_{\text{sample}}\leftarrow$Denoise(Sample, H, Hbn_1, Hbn_2)
5. $\text{detrend}_{\text{sample}}$, $\text{relative}_{\text{trends}}\leftarrow$Trend_Extraction($\text{denoise}_{\text{sample}}$, $\text{Season}_{\text{length}}$, Reg_1, Reg_2)
6. $\text{seasons}_{\text{tilda}}\leftarrow$Seasonality_Extraction($\text{detrend}_{\text{sample}}$, $\text{Season}_{\text{length}}$, H, K, Hbs_1, Hbs_2)
7. $\text{trends}_{\text{hat}}$, $\text{seasons}_{\text{hat}}$, $\text{remainders}_{\text{hat}}\leftarrow$Adjustment(Sample, $\text{relative}_{\text{trends}}$, $\text{seasons}_{\text{tilda}}$, $\text{Season}_{\text{length}}$)
8. if Trial $\neq$ 1 then
   converge$\leftarrow$Check_Converge_Criteria($\text{previous}_{\text{remainders}}$, $\text{remainders}_{\text{hat}}$)
   if converge=false then return false
   else return[input, $\text{trends}_{\text{hat}}$, $\text{seasons}_{\text{hat}}$, $\text{reminders}_{\text{hat}}$]
   end
9. Trial$\leftarrow$Trial+1
10. $\text{previous}_{\text{remainders}}\leftarrow\text{remainders}_{\text{hat}}$
11. Sample$\leftarrow\text{trends}_{\text{hat}}+\text{seasons}_{\text{hat}}+\text{reminders}_{\text{hat}}$
12. return[input, $\text{trends}_{\text{hat}}$, $\text{seasons}_{\text{hat}}$, $\text{reminders}_{\text{hat}}$]

由算法 1 可知,降噪步骤负责对每次迭代的序列数据进行预处理,通过减少高频噪音的数据占比提升提取步骤的稳健性,此阶段采用双边滤波(bilateral filtering)处理时序样本(Sample)(第 4 行),能够通过保留趋势突变和残差的 a_t,为强相关的 TSF 提供可解释性。分阶段提取是考虑到趋势和季节分量在变化频率上存在差异,若同步提取必然会影响到 TSF 的提取纯度。针对预处理后的迭代序列,趋势提取步骤通过 LAD 损失回归法分解到趋势分量(第 5 行),首先利用季节性差分操作降低季节性分量的影响系数,再通过对包含 LAD 的经验误差、趋势分量的一阶和二阶差分算子约束在内的三项加权目标函数进行最小化操作,在恢复趋势一阶差分的同时赋予趋势分量捕捉突变和对异常保持鲁棒性的能力,最后利用矩阵运算提取到本次迭代的趋势分量。基于分解相对趋势信号后的序列,季节提取步骤利用非局部季节性过滤法分解到季节分量(第 6 行),该方法对领域中数值点的权重赋值标准进行了优化,不完全依赖于在时间维度上的聚类属性,同时

考虑到领域中数值点在季节性上的相似程度，使得提取到的季节分量对异常值具备鲁棒性。特征调整步骤负责本次迭代过程中趋势和季节分量的数值调整(第 7 行)。通过交替地执行以上步骤，可以刷新每次迭代结果的估计值，在逼近收敛参数的同时提升最终 TSF 的准确性和鲁棒性(第 8 行)。

基于 RobustSTL 算法分解多元时间序列提取到的 TSF，理论上可以满足时间序列预测对数据准确性和鲁棒性的评价标准，能够通过深度学习算法实现序列演化规律和波动频率的无监督学习，在保证鲁棒性的同时提高预测准确性。考虑到 TSF 构建的特征二元组是将多元时间序列转换为二维向量组，纯粹地降维、降噪操作可能会导致因属性维度难以匹配模型复杂度而出现模型欠拟合情况，因此设置基于 TSF 和目标参数的相关性评估。相关性评估阶段通过运算相关性曲线的吻合度评判 TSF 在原始序列中的价值占比，衡量 TSF 是否具备弥补属性维度不足的强相关特性。本文采用皮尔逊相关系数(Pearson correlation coefficient)来运算 TSF 和预期目标参数之间线性相关的方向和强度。通常皮尔逊相关系数被定义为 $Pearson$，其计算式如下：

$$Pearson=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{4}$$

其中，X_i 分别赋值为趋势性特征集合和季节性特征集合，Y_i 赋值为目标参数集合。简单描述 $Pearson$，即为 X 和 Y 两个变量的协方差与其标准差乘积的比值，能够表示 TSF 集合和目标参数集合两组衡量定距变量间的线性关系。依据皮尔逊相关性系数对运算结果的评估，当 $pearson\in[0.8,1]$ 时，即证实 TSF 向量和预期目标参数之间具备极强的相关性，能够通过相关强度匹配深度学习算法的复杂程度。

2.2　基于特征再抽象(FRA)的数据驱动预测算法

本文通过遵循数据驱动的执行模式提出了基于 FRA 的数据驱动预测算法，能够针对诸如金融股票、医学脉冲信号等多元时间序列数据，通过序列型特征提取技术挖掘出历史数据中有价值的可辨识信息，再利用机器学习方法分析此类信息探知社会现象的发展历程和规律性，进而引伸外推以预测到未来时刻下事物的发展趋势，实现从已知事件测定未知事件的目的。

数据驱动的执行模式是通过结合特征提取和数据建模两项关键技术，处理以海量数据为基础的科技任务。本文通过融合 FRA 算法取代传统的特征选定机制，利用具有代表性且性能均衡的深度学习模型，在传统数据驱动策略的基础上实现优化。

基于多元时间序列数据普遍呈现出稳定上升、下降或者周期性波动的趋势依赖关系，导致基于线性多维数据提取到的趋势性和季节性特征(TSF)以及原始序列在变化规律和波动频率等图形特性上保持着极高的相似性，本文基于 TSF 向量构建多元时间序列预测的输入量，旨在通过相似的图形特性绘制出未来时刻的序列曲线，为多元时间序列预测的研究工作提供创新性的科研思路。

本文提出的 FRA 算法通过参数约束 TSF 向量的提取纯度，控制模型输入量和目标参数的相关强度以最大化数据的价值密度。该算法基于 RobustSTL 时间分解算法对多元时间序列进行属性拆解，通过属性降维的机制缓解模型针对多元数据集的处理负载，避免了深度学习模型的过拟合情况。RobustSTL 的递归过程中，通过迭代执行双边滤波和分解残差向量的操作实现数据降噪，保证了 TSF 向量的高价值密度。通过属性拆解后提取的 TSF 向量的提取纯度会随着迭代次数的增长逐渐提升，但为保证 TSF 向量的相关强度能够和深度学习的运算强度相匹配，FRA 算法设置了相关性评估步骤进一步核实。该算法基于皮尔逊相关系数运算 TSF 向量和目标参数之间的相关强度系数，以运算结果是否达到 80%来评估 TSF 向量能否作为模型的输入量。通过 FRA 算法提取到的 TSF 向量实现了多元数据向二维数据的转换，在属性降维的同时采用数据降噪和递归迭代的机制提升了特征的相关强度，在数据层面上提升了预测的准确性和鲁棒性。

本文选择 LSTM 网络作为数据建模的深度学习模型，该算法常被用于处理工程性的多元时间序列预测任务，具备稳定的信息继承和自学习能力。区别于 RNN 网络基于系统状态构建的递归计算模式，LSTM 网络通过设置输入门、遗忘门和输出门对独立的网络单元分别构建自循环，其中输入门主要承担当前时间步的系统输入和前一个时间步的系统状态对内部状态的更新操作，遗忘门主要承担前一个时间步内部状态对当前时间步内部状态的更新操作，输出门主要承担内部状态对系统状态的更新。LSTM 网络基于门控机制的优化措施，使得在处理长期依赖性问题时仍然可以避免梯度消失和梯度下降问题，在算法层面上提升了预测的准确性和鲁棒性。

3　实验与分析

本文采用的锂离子电池的充放电数据会随着使用时长的叠加，在宏观上呈现出持续性的损耗趋势，数据的波动频率会随着连续性的充电和放电机制呈现出周期性的起伏变化，明显的序列型特征和时间依赖特性使得此类数据能够作为实验的样例数据。另外，锂离子电池数据是和居民日常密切相关的数据，其数据规模和产出频率是相当可观的，基于此类数据得到的属性预测结果能够作为消防部门应对异常检测及安全预警的评估指标，也能作为工业厂商评估电池性能的考察标准，具有很强的社会效益。

3.1　数据集

实验阶段采用的锂离子电池数据集是由 NASA Ames 的 Prognostics CoE 提供，共计 5 类测试工况各不相同的庞大数据集。本文基于数据完整性、属性多样性等标准，最终选定了 7 组数据，其每组数据的规模均达到 50 000＋，详细数据集如表 1 所列。

表 1　实验相关数据集

Table 1　Experiment related datasets

Dataset：Battery Data Set	
Sets_Name	Battery_id
BatteryAgingARC_FY08Q4	05，07，08
BatteryAgingARC_25-44	33，34
BatteryAgingARC_45_46_47_48	46，47

上述数据集的差异是由于数据采集实验设置的环境变量和测试工况不同而导致的。实验选取的各数据集，其特征标签完全相同，为避免特征向量造成混淆，我们在数据准备阶段按照编号对数据集进行命名，命名规则为 Bid_discharge_soh。

以编号为 05 的数据集作为样例，其数据属性及属性说明如表 2 所列。

表 2 编号 05 数据集的详细数据

Table 2 Detailed data of No. 05 dataset

Dataset: B05_discharge_soh. csv	
Data_Type_Name	Meaning and Calculation_Formula
Voltage_terminal	Measured Voltage
Current_terminal	Measured Current
Temperature	Ambient Temperature(degree C)
Current_Charge	Current measured at charger(Amps)
Voltage_Charge	Voltage measured at charger(Volts)
time	Date and Time of the start of the Cycle
Capacity	Integrate the Current_Terminal
Cycle	one "discharge-charge" is one Cycle
Soh	The ratio of Measure Capacity to max Capacity

数据集的原始数据中，电池容量项(Capacity) 可以根据物理学的电流积分法计算得到，依据线性相关的特点直观地表征电池健康状态(SOH)。在实验阶段将电池容量项(Capacity)和电池健康状态项(SOH)设置为目标参数。

3.2 实验内容

本文采用基于数据驱动的大数据智能分析技术实现多元时间序列预测任务，包括特征再抽象和预测建模两个关键步骤。本节将预测任务的实验流程以数据流向图进行描述，依据数据格式分为 4 个阶段，如图 1 所示。

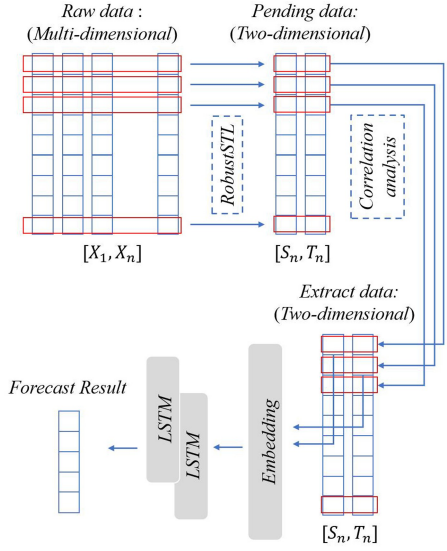


图 1 实验设计数据流向示意图

Fig. 1 Schematic diagram of data flow in experimental design

阶段 1 将作为模型输入的电池损耗数据描述为多维初始数据(Multidimensional Raw Data)。目前普遍被使用的序列型数据集几乎涵盖了相关应用领域的全部属性值，传统大数据模式下的序列型数据分析任务通常将该类数据集的标签项直接作为预测模型的输入向量集，导致因数据维度高而频繁出现过拟合问题，对数据分析结果造成影响。

针对上述频繁出现的情况，阶段 2 利用特征再抽象技术实现对多维初始数据特征执行数据降维处理，有效地解决因数据维度而产生的过拟合问题。该阶段通过将处理时间属性分离任务的 RobustSTL 算法作用于初始特征向量组，分解得到季节特征和趋势特征。多维度数据向二维特征向量(Two-dimensional Extract data)转化的过程，从根源上缓解了过拟合问题。

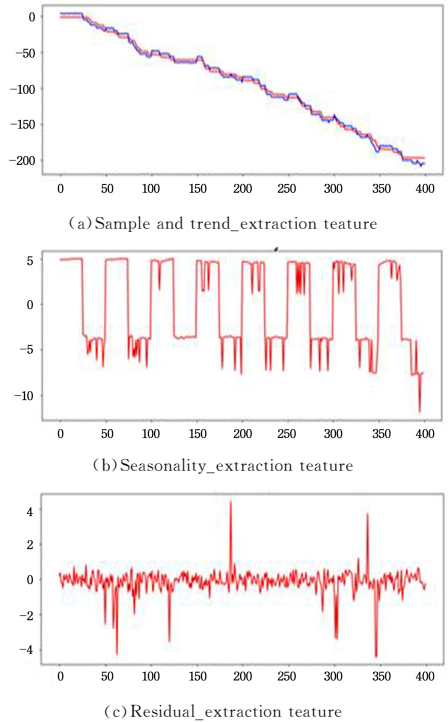
通过阶段 2 的降维处理得到低纬度特征向量，使得过拟合问题得以缓解，模型训练和预测的速度得以提升，但通过降维策略调整预测精度的方案对特征向量和目标参数的相关性提出了更高的要求。阶段 3 通过对提取的二维特征向量和目标参数执行相关性分析(Correlation analysis)，判断特征向量是否满足作为模型输入集的条件。

满足评价指标的高契合度特征通过嵌入层调整后进入最后阶段执行数据预测步骤。该阶段将具有信息继承能力的 LSTM 神经网络作为预测模型，通过预测结果对再抽象技术做出评价。

3.3 实验结果

图 2 给出了利用 RobustSTL 分解算法提取到的时域特征，包括趋势特征、季节特征和残差/噪音。图例中选取电池一个充放电周期(基本上以 400 数据周期变化)内的终端电压数据开展特征提取，沿横向坐标轴可以看到数据长度为 400，坐标轴纵向的跨度(最大值与最小值的差值)是根据特征提取强度和提取长度间隔决定的，不同特征具有不同的运算规则，例如趋势特征纵坐标轴的差值是提取强度和提取长度间隔的乘积。

图 2(a)展示了样本数据和趋势提取项，两条数据曲线基本重合，证实趋势性具有模拟数据演变特性的能力。图 2(b)展示了季节提取项，季节性通过平滑和高频波动来模拟数据不稳定变化，利用峰值和低谷的递归变化体现样本的周期性波动。图 2(c)展示了数据噪音，这些噪音是由 RobusSTL 分解算法通过调整噪音分离参数剔除得到的，以此降低数据采集阶段其他非实验因素的影响。



(a) Sample and trend_extraction teature

(b) Seasonality_extraction teature

(c) Residual_extraction teature

图 2 样本数据的特征提取图

Fig. 2 Feature extraction diagram of sample data

表 3 列出了将 7 组电池数据集的终端电流-电压数据抽取获得的时域特征序列与 8 组数据集对应的目标参数(Capacity，SOH)作为皮尔逊相关系数计算公式的输入序列所获得的相关性评价数值。

表 3　终端电流-电压提取特征相关性分析

Table 3　Correlation analysis of terminal current-voltage extraction features

（单位：%）

DataSets_Name	Trend_val	Seasonality_val
B05_discharge_soh	91.46	87.12
B07_discharge_soh	92.14	89.28
B08_discharge_soh	93.78	91.41
B33_discharge_soh	92.03	88.16
B34_discharge_soh	93.88	90.59
B46_discharge_soh	89.63	88.37
B47_discharge_soh	90.20	86.90

由表 3 可知，Trend_val 列的数值基本都达到 90%以上，表明由终端电流-电压所提取的趋势性特征和目标参数之间具备极强的相关性，证实电流或电压的变化情况可以反映电池的健康状态，该实验现象也和实际场景下电池蓄电状态的情况一致。

Seasonality_val 列的数值大小体现了目标参数在波动性和周期性方面与初始数据的相关性程度，虽然数值基本保持在 90%以下，并未达到极强相关性的评价标准，但基本趋近于强相关性的评价上限，在实验模型预测阶段被认为满足作为模型输入向量的标准。

实验针对 7 组样本数据执行 FRA 算法，并基于 Init_f 列（模型输入为初始特征集合）、Ext_f 列（模型输入为提取特征集合）和 Ext&Init_f 列形成对照，利用长短期递归神经网络（LSTM）执行建模预测的结果如表 4 所列。

表 4　长短期递归神经网络（LSTM）预测结果

Table 4　Longshort-term recurrent neural network(LSTM) prediction results

（单位：%）

Sets_id	Init_f	Ext_f	Ext_Init_f
_05	86.35	89.88	88.25
_07	85.31	89.67	90.01
_08	84.76	87.92	88.27
_33	86.67	89.71	87.35
_34	84.29	88.68	88.12
_46	85.73	88.35	89.94
_47	82.99	86.17	85.66

对比 Init_f 列和 Ext_f 列数值可见，利用特征再抽取技术获得的季节趋势特征完成数据预测任务在效果上优于传统数据驱动预测方法。鉴于数据规模庞大，我们将造成该现象的因素大致归结为：1）初始数据集的维度达到 9，满足高维度特征集标准，可能出现超出 LSTM 预测模型的计算能力，导致过拟合问题，使得 Init_f 列效果不理想；2）数据来源于电池充放电的测定数值，可能存在平滑处理方法难以剔除的数据噪音，对模型预测效果造成了干扰。

表 4 的 Ext&Init_f 列是将电池数据集的终端电流-电压特征和时域特征集成后作为输入集得到的预测结果。对比 Init_f 列，发现数值明显优于 Init_f 列，证实了过拟合因素造成预测效果不理想的猜想。对比 Ext_f 列，发现数值差距不明显，在 8 组数据中仅有 3 组优于 Ext_f 列，说明具备反应数据演变趋势和波动变化的时域特征可以更优于初始数据，该现象得益于特征提取阶段剔除了噪音数据。

结束语　针对基于强依赖于时间属性的高维度序列型数据预测任务，提出了基于特征再抽象（FRA）的多元时间序列预测方法。此预测方法是基于数据驱动的运行策略，通过学习历史观测数据的序列型特征来建模历史数据和未来数据之间潜在的映射关系；数据准备阶段采用 FRA 算法完成原始序列趋势性和季节性特征（TSF）的二阶抽象提取，实现由多元时间序列标签属性特征向二元特征向量组的转换，并结合相关性评估算法衡量 TSF 向量组和预期目标参数间的关联强度，旨在在数据基础上保障预测的准确性和鲁棒性；数据建模阶段借助长短期记忆网络（LSTM）完成预测任务，并依据性能指标评价 TSF 映射序列关键信息的能力，验证了 FRA 算法能够通过构建鲁棒且准确的数据基础提升整体预测效果。

本文提出了一种多元时间序列预测方法，其创新点一是提出了基于先抽象后评估递进式机制的 FRA 算法，首先将高维度原始标签特征集转换为低维度的 TSF 向量组，通过属性降维和数据降噪的机制避免了深度学习模型出现过拟合情况，再通过相关性评估算法确保抽取的 TSF 向量组能够凭借极强的关联性弥补数据维度和模型架构间的差异，减少深度学习模型出现欠拟合情况；创新点二是提出了将 FRA 算法内嵌于特征提取步骤的数据驱动预测方法，通过融合 FRA 算法构建低维度、强相关和高价值密度的模型输入量，再结合深度学习算法强大的运算能力保障预测的准确性和鲁棒性。本文基于 7 组真实数据集进行对比实验，结果表明，基于 FRA 的多元时间序列预测算法具有较好的整体预测性能，面向海量高维度的序列集时其预测性能的表现更加卓越。

当前工作主要依托实际预测任务验证 TSF 等时域特征能否替代原始数据作为模型的训练向量集，未来工作可以继续探究时域特征可能的应用场景或学术领域。

参 考 文 献

[1] REN S G，ZHANG J X，GU X J，et al. Overview of Feature Extraction Algorithms for Time Series[J]. Journal of Chinese Computer Systems，2021，42(2)：271-278.

[2] ZHAO D F，HUANG Y L，HUANG D M，et al. Research on time series motif association rule mining method based on AR_TSM[J]. Application Research of Computers，2021，38(2)：403-408.

[3] YANG H，WANG H Q，CHENG D J. Series Outlier Data Mining Based on Forecastment[J]. Computer Science，2004(4)：117-119，146.

[4] YE L，KEOGH E. Time series shapelets：a novel technique that allows accurate，interpretable and fast classification[J]. Data Mining and Knowledge Discovery，2011，22：149-182.

[5] WAN C，LI W Z，DING W X，et al. A Multivariate Time Series Forecasting Algorithm Based on Self-Evolution and Pre-training[J]. Chinese Journal of Computers，2022，45(3)：513-525.

[6] JIA J，HU X S，DENG Z W，et al. Data-driven Comprehensive Evaluation of Lithium-ion Battery State of Health and Abnormal Battery Screening[J]. Journal of Mechanical Engineering，2021(3)：87-97，57.

[7] SHI X，CHEN Z，WANG H，et al. Convolutional LSTM Network：A Machine Learning Approach for Precipitation Nowcasting[J]. arXiv：1506.04214，2015.

[8] VASWANI A. Attention is All You Need [J]. arXiv：1706.03762，2017.

[9] ELHASSAN T A M，RAHIM M S M，SWEE T T，et al. eature

Extraction of White Blood Cells Using CMYK-Moment Localization and Deep Learning in Acute Myeloid Leukemia Blood Smear Microscopic Images[C]// IEEE Access. 2022: 16577-16591.

[10] LIU L, ZHU J C, HAN G J, et al. Bearing health monitoring and fault diagnosis based on joint feature extraction in one-dimensional convolution neural network[J]. Ruan Jian Xue Bao/Journal of Software, 2021, 32(8): 2379-2390.

[11] MA C C, DU X H, CAO L F, et al. Burst-Analysis Website Fingerprinting Attack Based on Deep Neural Network[J]. Journal of Computer Research and Development, 2020, 57(4): 746-766.

[12] ZOU X Y. TIme series prediction algorithm based on graph laplace transform and extreme learning machine[J]. Computer Applications and Software, 2021, 38(4): 288-294.

[13] GUO Y H, LU J Y, HUANG C H, et al. Mesh Texture Smoothing Based on Hybrid Spectral Encoding[J]. Chinese Journal of Computers, 2021, 44(2): 318-333.

[14] JIA Z Y, LIN Y F, LIU T H, et al. Motor Imagery Classification Based on Multiscale Feature Extraction and Squeeze-Excitation Model[J]. Journal of Computer Research and Development, 2020, 57(12): 2481-2489.

[15] LIU Y Y, LI J P, BAI H F, et al. Trend feature extraction method for time series based on turning point and trend segment [J]. Journal of Computer Applications, 2020, 40(S1): 92-97.

[16] ZHOU Q, WU T J. Trend feature extraction method based on important points in time series[J]. Journal of Zhejiang University(Engineering Science), 2007(11): 1782-1787.

[17] WIJSEN J. Trends in databases: reasoning and mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(3): 426-438.

[18] WEN Q, GAO J, SONG X, et al. RobustSTL: A Robust Seasonal-Trend Decomposition Algorithm for Long Time Series[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 5409-5416.

[19] CLEVELAND R B, CLEVELAND W S. STL: A seasonal-trend decomposition procedure based on Loess[J]. Journal of official statistics, 1990, 6(1): 3-73.

WANG Hao, born in 1998, postgraduate. His main research interests include big data mining and intelligent analysis technology.

ZHOU Jiantao, born in 1974, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include cloud computing and software engineering.