基于深度学习的大数据时序预测模型构建与分析

徐瀚杰* 田乐乐 胥立军 肖凌俊
广州新华学院 广东广州 510000

摘 要：时序预测是大数据分析领域的核心任务之一，其核心是通过挖掘时序数据的内在规律，实现对未来状态的精准推断。针对大数据环境下时序数据规模庞大、特征复杂、动态性强等问题，传统预测方法难以满足预测精度与效率需求。本文基于深度学习技术，构建一套完整的大数据时序预测模型，系统分析时序预测基础理论、模型构建框架、数据处理方法、训练优化策略及性能评估体系，探讨深度学习在时序预测中的适用性与优势，解决大数据时序预测中的关键技术难题，为相关领域的时序预测应用提供理论支撑与方法参考，全文围绕模型构建与分析展开，兼顾理论严谨性与实践指导性。

关键词：深度学习；大数据；时序预测；模型构建；性能评估

引言

大数据技术极速发展之际，金融、交通、环境等诸多领域不断产生时序数据，其中蕴含的时间关联信息对于决策颇具价值。时序预测是发掘时序数据价值的关键所在，因而备受学界和业界关注。传统时序预测方法需依靠人工特征获取，很难应对大数据环境下时序数据具有的高维数、非线性以及动态特性等状况，其预测准确度和泛化能力存在局限性。深度学习技术具备较强的特征自动学习及非线性拟合能力，给解决大数据时序预测问题带来新途径。本文围绕基于深度学习的大数据时序预测模型构建与分析展开研究，梳理时序预测基础，构建模型框架，优化训练策略，评估模型性能，旨在提升大数据时序预测的精度与效率，推动深度学习在时序预测领域的应用与发展。

一、时序预测基础与问题分析

（一）时序数据特性与预测任务定义

时序数据指的是依照时间先后次序形成起来的一系列数据点的集合，它具备时间关联性、动态性、趋势性以及周期性的特点。时间关联性属于时序数据的基本属性，相邻的数据点彼此间有着密切的依存关系，此依存关系成为时序预测的关键。所谓动态性就是说时序数据会随着时间的流逝而发生动态改变，这由许多潜藏要素所影响并引发波动现象。所谓趋势性就是显示数据在长时间跨度下表现出上升、下降或者稳定这样一种变动规则。周期性意味着数据会在固定的时段区间里反复再现类似的模式。做时序预测的时候，重点在于依靠过去已有的时序数据，并采用某种建模手段来找出数据内部蕴含的时间联系规则，从而对将来某个确切时刻或者某段时间内的数据数值执行推算和估算操作。按照预测时限长短可以区分为短期、中期以及长期三种类型。

（二）大数据环境下时序预测的挑战

处于大数据环境当中，时序预测碰上诸多明显难题。其一是数据量大造成的处理负担沉重，如此众多的时序数据要完成存储、输送以及运算，必然耗费不少资源，以往的处理手段往往无法达到即时性的要求。其二，大数据中的时序数据常常具有高维、噪音、不完善等特性，多源数据存在异构情况，使得特征获取和建模更为困难，极易引发预测出现偏差。其三，时序数据的动态性质和非线性特点越发突出，数据分布不断随时间而改变，以往的模型缺乏适应能力，难以把握数据复杂的变动规则。大数据时序预测对于即时性有着较高的要求，如何既能保障预测精度又能够优化模型计算效率以达成快速预测是当下遭遇的关键难点所在。

二、模型构建理论框架

（一）深度学习时序建模理论基础

深度学习做时序建模的时候，核心理论依靠神经网络的特征学习以及时间依赖建模。这个核心想法就是通过多层非线性变换，把时序数据投影到高维特征空间当

课题项目：本文得到多模态大模型赋能课程与教学的改革与实践项目（项目编号：2025J052）的支持。

通讯作者简介：徐瀚杰，高级工程师，硕士，研究方向为机器学习、大模型及数据科学。

中，从而找出数据之间的时间关联规则。时序建模包含几大核心理论，时间依赖建模理论、特征层级学习理论以及端到端学习理论。时间依赖建模理论重点在于捕捉时序数据里相邻数据点之间的依赖关系，它靠循环结构或者注意力机制来有效地利用历史数据信息；特征层级学习理论看重的是借助多层网络慢慢获取数据从表层到深层的各种特征，做到对复杂特征细致入微的描绘；而端到端学习理论达成了从原始时序数据到预测结果的直接转换，减少了人工特征干预的必要，优化了建模速度和预测的一致性，给模型结构设计提供了关键的理论支持。

（二）模型结构设计与组件选择

深度学习的大数据时序预测模型结构设计要依照时序数据特性和预测需求来执行，其核心部件包含输入层、特征获取层、时序建模层以及输出层。输入层的任务是接收经预处理的时序数据，并把它转化为模型能识别的向量形式，从而满足后面特征获取和建模的需求。特征获取层利用卷积神经网络之类的结构，做到对时序数据浅层特征的获取并执行降维操作，去除多余信息。时序建模层属于模型的关键部分，可以选择循环神经网络或者注意力机制网络，以此抓住时序数据中的时间依赖关系，找出深层次的时间联系规则。输出层按照预测任务的要求，运用合理的激活函数，给出预测结果。组件的选择要兼顾模型性能和计算效率，使得模型符合大数据时序预测的需求，达成精度与效率的兼顾。

（三）损失函数与评价指标体系

损失函数是模型训练的关键依照，其用以度量模型预测成果与真实值之间的差异，引导模型参数的优化。对于大数据时序预测任务而言，要选取契合时序数据特性的损失函数，既要关注预测精度，也要重视模型的稳定性。常见的损失函数有均方误差损失、平均绝对误差损失等，可以遵照预测任务的重点实施选择和调节。评价指标体系可全方位评估模型预测表现，其中应包含预测精度、鲁棒性以及泛化能力等诸多层面。关键评价指标涉及平均绝对误差、均方根误差、决定系数等，它们分别用以度量预测结果的偏离程度和拟合状况。并且，按照大数据场景的需求，再加上计算效率相关的指标，从而创建起全面而科学的评价指标体系，给模型性能评定提供参考。

三、数据处理与特征工程

（一）大数据预处理与质量控制

大数据预处理是形成时序预测模型的先决条件，关键目的在于提升数据质量，给后续的特征工程和模型训练赋予可靠的数据支持。预处理流程包含数据清洗、数据标准化和数据补全这三大阶段。数据清洗要清除时序数据里的噪音、异常值以及多余数据。数据标准化可缩减不同维度数据之间的量纲差别，并把数据归结到相同的范围内，从而加快模型训练的速度并加强其稳定性。而数据补全是应对时序数据中出现的空缺值情况，遵照数据的时间相关特性去填充缺少的信息，以此保证数据的完整性，进而为后续的特征获取和建模创建前提。

（二）时序特征提取与表示学习

时序特征获取与显示学习对于优化模型预测精度十分关键。依靠深度学习的特征获取不需要人工去设计特征，而是由神经网络自行从时序数据里学得深层次的特征，涵盖时间域特征和频率域特征。时间域特征着重关注数据的方向、规律性和相关性，而频率域特征则用来探寻数据潜藏的波动规律。显示学习通过特征映射，把原始的时序数据变成低维且紧凑的特征向量，这样既能保存住数据的主要信息，又能缩减数据的维度，减小模型的计算量，而且还能加强特征的可解读性，给模型训练提供优质的特征来源。

（三）数据切片与样本生成策略

数据切片和样本生成对于适配深度学习模型训练十分关键，其核心在于把连续的时序数据转为成模型可用以训练的样本数据。数据切片依照时间窗口策略，按照预测任务所需的时间范围来设定合理的窗口大小，从而把连续的时序数据分割成许多个彼此相关联的子序列，各个子序列包含一段固定长度的历史数据及其对应的预测目标值。在生成样本的时候，要妥善划分训练集、验证集以及测试集，保证样本具有随机性与代表性。通过对时序数据实施合理的变换，扩充样本的数量，优化模型的泛化能力，进而满足大数据环境里模型训练的需求，使得模型得以深入把握时序数据的核心规律。

四、模型训练与优化方法

（一）分布式训练架构设计

在大数据时序预测当中存在数据规模巨大的情况，而分布式训练架构的设计对于优化模型训练效率十分关键。该架构依靠集群计算理论，把大量的时序数据以及模型训练任务分配给众多节点以执行并行计算，从而缩减训练所需的时间。模型包含数据节点、计算节点以及参数服务器节点。数据节点承担数据的储存和传送工作，计算节点专门执行局部模型训练，参数服务器节点则负责整合和更新全局参数。这些节点相互协作，保障分布式训

练既稳定又一致，进而满足海量时序数据训练的需求。

（二）超参数优化与正则化技术

超参数优化和正则化技术对于提升模型性能、防止过拟合十分关键。超参数包含学习率、批次大小、网络层数等，这些参数的取值会直接左右模型的训练成果以及收敛快慢。超参数优化通过合理的搜索策略来找出最理想的超参数集合，从而协调好模型训练的精准度和收敛的高效性，规避因为超参数设定不合理而引发的模型性能下滑情况出现。正则化技术可用来解决模型过拟合的问题，它会在损失函数里面加上正则项，以此来约束模型参数的复杂程度，减小模型对于训练数据的过度拟合现象，进而加强模型的泛化能力。常见的正则化手段有L1、L2以及Dropout技术。

五、预测性能评估与验证

（一）多尺度预测精度对比

多尺度预测精度对比属于评估模型预测性能的关键部分，目的在于从各类时间跨度角度出发，全方位检测模型的预测能力。按照预测时间跨度分别计算各个尺度对应的评价指标，然后比较模型在这些尺度上的预测精度。拿该模型和传统的时序预测方法以及基本的深度学习模型做对比，探究所创建模型在各个尺度存在的长处和短板。借助误差分析找出模型预测出现偏差的主要原因，进而给模型的优化指示方向，使得模型在各种不同的预测情形之下都能取得较为理想的预测成果。

（二）模型鲁棒性与泛化能力分析

模型的鲁棒性与泛化能力属于衡量模型实用性的关键指标，其中鲁棒性体现的是模型在遭遇噪声、异常值或者数据分布改变时所表现出的稳定性，而泛化能力则关乎到模型对于未见过的测试数据实施预测的能力。执行鲁棒性分析的时候，可以给测试数据增添一些合理的噪声，然后查看模型预测结果出现的变化范围，以此来评判模型抵御干扰的能力。至于泛化能力分析，则要拿模型在训练集和测试集上的表现做比较，再配合交叉验证法，看下模型是不是存在过拟合或者欠拟合的情况，从而保证模型可以适应各种场景中的时序数据，解决大数据环境里数据不断变动的问题。

（三）计算效率与可扩展性评估

计算效率和可扩展性这两方面需契合大数据时序预测的实际应用需求，重点在于评估模型的训练速度、预测速度以及其针对数据规模的适配能力。计算效率的评估通过统计模型的训练历时、预测历时来展开，分析模型在不同数据规模时的计算耗时情况。而可扩展性的评估则是逐步扩大数据规模，留意模型性能和计算效率的变动趋向，从而判定该模型是否具备应对海量时序数据处理需求的能力。经过对计算效率及可扩展性的评估，可以验证模型在大数据环境下的实用价值，保证模型能够达到实际应用中即时预测和大规模数据处理的需求。

结语

本文就依托深度学习创建大数据时序预测模型并加以分析展开研究，梳理了时序预测的基本原理，创建起较为完善的模型理论架构，并论述了数据处理、模型训练优化以及性能评定这些关键方法。研究表明，深度学习能够很好地适应大数据时序数据的特点，借助自动识别特征并形成时间关联的机制，明显提升了预测的准确度和速度。本文所创建的模型在多尺度预测能力、稳定性以及计算速度方面均表现不错。未来还可以进一步优化模型的内部构造，加强其解释能力，从而扩展其应用范围，给大数据时序预测提供更为全面的理论及方法支持。

参考文献

[1]Surur F M, Mamo A A, Gebresilassie B G, et al. Unlocking the power of machine learning in big data: a scoping survey[J]. Data Science and Management, 2025,8(4):519-535.

[2]严冬梅，杜锦华，王烁杰.引入凸包知识蒸馏技术的时间序列预测方法[J].数据分析与知识发现，2025，9（10）：160-175.

[3]Chen Z, Lin R, Xu B, et al. LXformer: A Long-Term Time Series Forecasting Model Based on Multi-Granularity Feature Extraction for Edge Devices[J]. Big Data Mining and Analytics,2026.

[4]丛国庆.基于时序负载预测的弹性大数据服务系统设计与实现[D].山东大学，2024.

[5]马超红，郝新丽，孟小峰，等.机器学习赋能的多维数据查询处理研究综述[J].计算机学报,2025,48（1）.

[6]苏丽，孙雨鑫，苑守正.基于深度学习的实例分割研究综述[J].智能系统学报，2022，17（1）：16-31.