微调长短时记忆网络群：一种小样本多变量时序预测算法

鲍观明1

1(南京信息工程大学 计算机学院、网络空间安全学院, 江苏 南京 210044)

摘 要: 跨项目软件缺陷预测技术可以利用现有的已标注缺陷数据集对新的无标记项目进行预测,但需要两者之间具有相同的度量集合,难以用于实际开发.异构缺陷预测技术可以在具有异构度量集合的项目间进行缺陷预测,该技术引起了大量研究人员的关注.现有的异构缺陷预测技术利用朴素的或者传统机器学习方法为源项目和目标项目学习特征表示,所学习到的特征表示能力很弱且缺陷预测性能很差.鉴于深度神经网络强大的特征抽取和表示能力,基于变分自编码器技术提出了一种面向异构缺陷预测的特征表示方法.该模型结合了变分自编码器和最大均值差异距离,能够有效地学习源项目和目标项目的共性特征表示,基于该特征表示可以训练出有效的缺陷预测模型.在多组缺陷数据集上通过与传统跨项目缺陷预测方法及异构缺陷预测方法实验对比验证了所提方法的有效性.

关键词: 异构缺陷预测;变分自编码器;特征表示

Fine-Tuning Group LSTM (FTG-LSTM): A Novel Approach for Generalized Multivariate Time Prediction with Few Samples Problems

Bao Guanming1

1(School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China)

**Abstract**: Cross-project defect prediction technology can use the existing labeled defect data to predict new unlabeled data, but it needs to have the same metric features for two projects, which is difficult to be applied in actual development. Heterogeneous defect prediction can perform prediction without requiring the source and target project to have the same set of metrics and thus has attracted great interest. Existing heterogeneous defect prediction models use naive or traditional machine learning methods to learn feature representations between source and target projects, and perform prediction based on it. The feature representation learned by previous studies is weak, causing poor performance in predicting defect-prone instances. In view of the powerful feature extraction and representation capabilities of deep neural networks, this study proposes a feature representation method for heterogeneous defect prediction based on variational autoencoders. By combining the variational autoencoder and maximum mean discrepancy, this method can effectively learn the common feature representation of the source and target projects. Then, an effective defect prediction model can be trained based on it. The validity of the proposed method is verified by comparing it with traditional cross-project defect prediction methods and heterogeneous defect prediction methods on various datasets.

**Key words**: heterogeneous defect prediction; variational autoencoders; feature representation

# 引言

1.1 时间序列预测问题概述

时间序列预测是许多领域中的一个重要问题，包括工业和商业等应用。例如，公共交通运营商可以根据人流密度的预测，在特定地区合理部署运力以减小人群排队时间，从而提高运营收益。另一个例子是大型零售超市，如果能够准确预测不同商品在不同分区的销售额，就可以避免由于供应不合理而导致的收益损失。然而，现实世界中的时间序列通常表现出复杂的动态特性和不稳定性，这给时间序列预测任务带来了挑战。

现实世界的时间序列往往受到多种因素的影响，例如交通流量受到位置、天气和时间点等因素的影响，销售量受到节假日、商品售价、地区特性和气候变化等因素的影响。这些难以捕捉的动态特性使得时间序列预测任务变得复杂。在本文中，我们研究小样本多变量时序预测问题，即存在多个时序特征的时间序列预测任务，但序列样本数量有限。

传统的时间序列预测算法，如自回归积分移动平均模型（ARIMA）及其算法族、循环神经网络（RNN）等，即使在小样本短时任务中也难以表现出优秀的性能，特别是当时序特征稀疏且灵活时。近年来，随着深度学习的快速发展，相较于小样本的短时时序预测任务，大样本长序列时序预测任务得到了令人满意的解决。例如，在自然语言处理领域，Transformer模型通过使用多头注意力机制捕捉多个时序变量之间的相互关系，图卷积神经网络（GCN）将多变量时序数据表示为图结构，并利用图卷积操作来捕捉变量之间的相互关系。这些新兴的深度学习模型在海量样本的长时序预测任务中表现出色。然而，少量的数据难以让深度学习模型得到良好的训练，因此这些深度学习模型在小样本短时时序预测任务中的表现平平。

1.2 小样本多变量时序预测解决方法

1.2.1 传统方法

在小样本多变量时序预测问题中，传统的时间序列预测方法被广泛研究和应用。其中最常见的方法是自回归积分移动平均模型（ARIMA）及其算法族。ARIMA模型是一种基于线性统计模型的方法，通过对时序数据的自回归（AR）部分、差分（I）部分和移动平均（MA）部分建模，捕捉数据中的趋势和周期性。ARIMA模型可以根据数据的性质进行参数调整，以适应不同的时序模式。此外，还有一些基于统计分析和时间序列分解的方法，如季节性分解、指数平滑法等。季节性分解方法将时序数据拆分为趋势、季节和残差部分，以便更好地分析和预测其中的模式。指数平滑法则基于移动平均的思想，通过对数据进行平滑处理，提取其趋势和季节成分。然而，传统方法在处理复杂、非线性和高维数据时可能受到限制。这些方法通常依赖于对数据进行线性或统计性假设，并且对于具有长期依赖或非线性关系的数据，其预测能力可能不足。此外，传统方法在选择模型结构和参数时，需要依赖领域专家的经验和领域知识，因此在应对新的时序预测问题时可能缺乏灵活性和适应性。

1.2.2 深度学习方法

随着深度学习的发展，针对小样本多变量时序预测问题，研究人员提出了一些基于深度学习的方法。深度学习方法通过利用神经网络的强大表示学习能力和非线性建模能力，能够更好地处理复杂和高维时序数据。其中，循环神经网络（RNN）及其变种，如长短期记忆网络（LSTM）和门控循环单元（GRU），被广泛应用于建模时序依赖关系。RNN通过在网络中引入循环连接，可以捕捉到数据中的时间依赖关系。LSTM和GRU是RNN的改进版本，通过门控机制来控制信息的传递和遗忘，从而有效地解决了梯度消失和梯度爆炸等问题，更好地捕捉长期依赖。除了RNN，卷积神经网络（CNN）也被用于多变量时序数据的建模和特征提取。CNN通过卷积操作可有效地提取局部特征，并通过池化操作进行空间降维，使得网络能够更好地处理时序数据。在多变量时序预测中，可以将不同变量视为输入的不同通道，利用CNN对每个通道进行特征提取，并通过全连接层进行融合和预测。此外，注意力机制在多变量时序预测中也得到广泛应用。注意力机制可以用于捕捉不同变量之间的重要关系和注意力权重，从而在建模过程中对不同变量赋予不同的重要性。通过引入注意力机制，网络可以自动学习变量之间的相关性和重要性，提升预测性能。

1.3 本文研究内容及贡献

作为《机器学习》的课程论文，本文主要是利用机器学习课程中所涉及的深度学习方法对于多变量时间序列预测任务进行求解的研究与实践。本文的主要研究内容及贡献可以总结如下：

1. 首先，对多变量时序预测任务进行建模描述，并使用经典机器学习算法对其进行求解；
2. 接着，使用主流LSTM网络对多变量时序预测任务进行求解，并对比经典机器学习算法；
3. 最后，提出了一种微调的长短时记忆网络群算法，以更优的方式解决了短时多属性时序预测任务，并通过大量实验验证了算法的有效性。

# 基础知识

## 一般多属性时序预测问题定义：

在这一小节，我们讨论一般多属性时序预测问题的定义，我们给定一般的多属性时序预测问题定义如下：

|  |
| --- |
| **一般多属性时序预测问题定义**  **（i） 输入：**  **（ii） 输入描述：**  给定序列集合 ，其中 是序列样本数量； ，其中 是样本 的特征特维数，假设且，即每个样本特征个数相同；，其中 是每个样本的每个特征序列的长度。  **（iii）输出：**  有特征 ，给定，输出为：。 |

特别地，当 取值较小，且 时，问题演化为小样本多属性时序预测问题。

# 经典LSTM网络解决多变量时序预测问题

## 经典LSTM算法概述

LSTM（Long Short-Term Memory）是一种经典的循环神经网络（RNN）变体，用于处理和建模序列数据。相比于传统的RNN，LSTM网络在处理长期依赖性问题上表现更出色。LSTM网络的核心思想是引入了称为"门"的结构，这些门可以选择性地控制信息的流动和记忆的更新。如图1所示，LSTM网络的关键组成部分包括记忆单元，图中标志为A的整体，和三个门：输入门、遗忘门和输出门。

输入门：控制新输入信息对记忆单元的影响。它通过一个sigmoid激活函数来决定是否接受新的输入，并使用另一个tanh激活函数来生成一个新的候选值，用于更新记忆单元的内容。

遗忘门：控制之前的记忆是否被遗忘。它通过一个sigmoid激活函数来决定是否丢弃之前的记忆，以及保留多少之前的记忆。

输出门：控制输出的记忆单元的内容。它通过一个sigmoid激活函数来确定输出的哪些部分应该被输出，然后将其传递给一个tanh激活函数进行处理，得到最终的输出。

LSTM网络通过这些门的组合来实现对序列数据的建模和预测。通过输入门、遗忘门和输出门的控制，LSTM网络可以自动选择性地记住和忘记过去的信息，并从中提取有用的特征和模式。这使得LSTM网络够更好地捕捉长期依赖关系，避免梯度消失或爆炸的问题，从而在处理和预测序列数据时表现更出色。

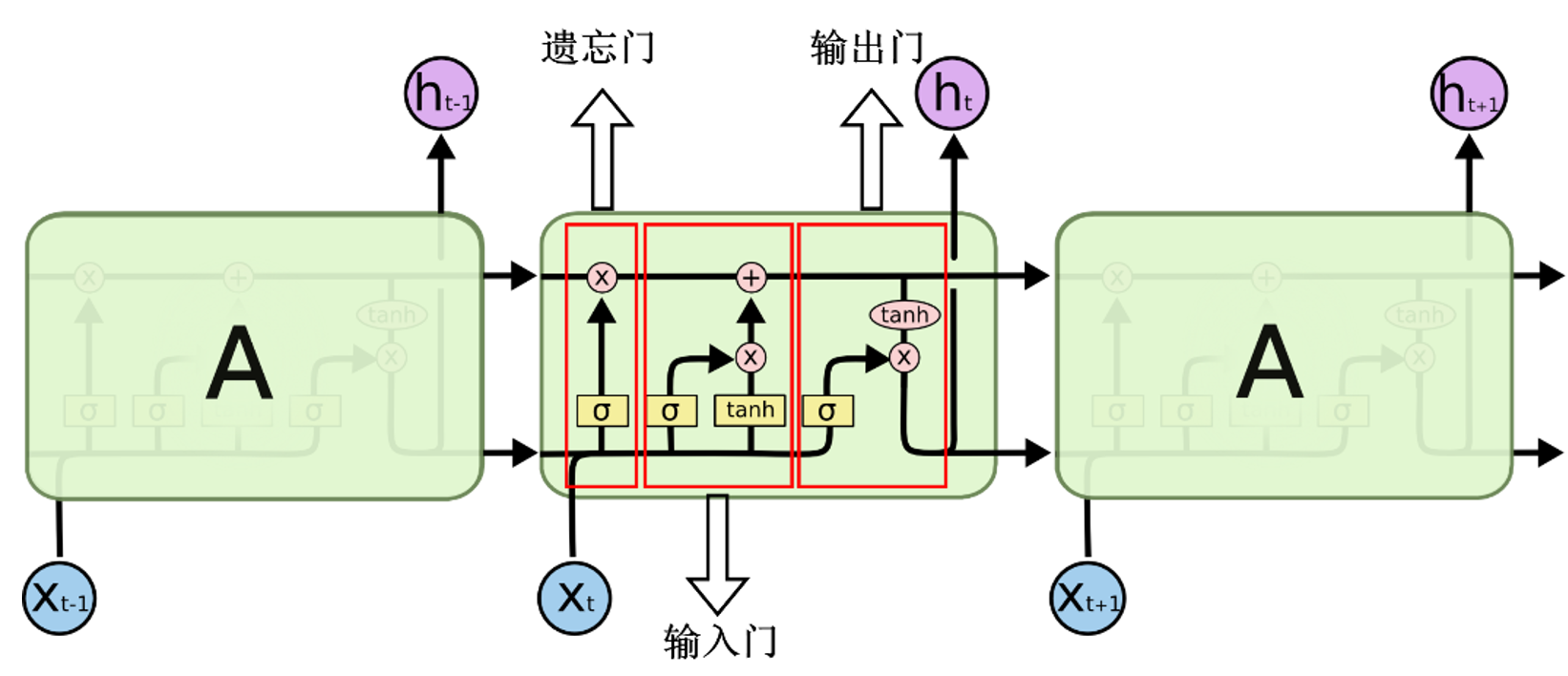


图1 经典LSTM结构

## 经典LSTM解决多变量时序预测问题

应用经典LSTM解决多变量时序预测问题的主要步骤在于：设计LSTM网络结构、损失函数选择、LSTM网络训练。

# 微调长短时记忆网络群算法（FTG-LSTM）

|  |
| --- |
| 算法1.微调长短时记忆网络群算法（FTG-LSTM）伪代码 |
| 输入：小样本多属性时序数据；微调权重；用于微调的比例；欠拟合度量；序列相似性指标  输出：   1. ； 2. **For** **do** 4. （若干epochs：前向传播，计算输出，计算损失，反向传播，参数更新） 6. **End** 7. **For do** 8. **For**  **do**  11. **End** 12. **End** 13. **For in Group do** 14. **If Accuracy(model, TrainSet) < do** 15. **TopNets**  Group内与model训练集相似性最大的()个model 16. **NetWeights** 17. **Model** |

# 实验分析

# 总 结

References:

1. Hall T, Beecham S, Bowes D, Gray D, Counsell S. A systematic literature review on fault prediction performance in software engineering. IEEE Trans. on Software Engineering, 2011,38(6):1276–1304.