DSTLinear: Decompostion Spatial-Temporal Linear

鲁权锋 201830168 [1]

Abstract

多元时间序列预测是时间序列分析领域的一个重要的研究方向，如何综合考虑多个变量之间的关系是多元时间序列分析中主要需要解决的问题。目前，有实验发现一个简单的线性模型（LTSF-Linear）就可以达到非常好的预测效果。但该模型将多元时间序列看作多个一元时间序列，多个维度共享同一个权重，忽略了变量之间的关联。对此，我在他们提出的 DLinear 模型做出改进，对趋势项加入时间和空间嵌入，综合不同维度、不同时间跨度上的信息，以求达到更好的预测效果。

1. Introduction

时间序列分析并不是一个新兴的领域，早在几百年前就有了很多关于该领域的研究，但由于它本身的不确定性和复杂性，时至今日，它依旧是一个值得深入研究和分析的方向。多元时间序列（Multivariate time series，MTS）是一种在多个维度上综合了时间和空间不同信息的时间序列数据，而长时间尺度的多元时间序列预测在非常多的场景下都扮演着重要的角色，一直是学界和业界的一个重要研究方向。如何表示数据、如何挖掘和综合不同维度和时间跨度的信息、如何降低模型的复杂度是多元时间序列预测任务的挑战和困难之处。

2. Relative Work

大体来说，时间序列分析，尤其在预测任务上一般有两种方法：统计方法和机器学习的方法。

统计方法主要是通过分析变量自己和不同变量间的相关性，建立统计模型，例如非常经典的 ARIMA。它其中的一种变体：VARIMA 是用于多元时间序列预测的。它主要通过计算不同维度之间的协相关系数来建模和刻画不同变量之间的相关性，因此在维度较高的情况下，它不可避免地带来极大的时空复杂度，这也成为了此类传统统计方法的瓶颈。

而机器学习方法主要是通过建立深度学习模型来刻画时序数据时空方面的信息，近年来提出了许多新颖而有效的方法。RNN 是一种经典的机器学习的模型，它本身不断迭代迭代的网络结构使其很天然地适用到时间序列预测任务上，已有实验已充分验证了它不错的效果。TCN(Bai et al., 2018) 是 CNN 的一个改进，它通过膨胀卷积和因果卷积机制，扩大了感受野，更关注了时间信息的提取，使其在时间序列分析上可以取得比 RNN 更好的效果。此外，在自然语言处理（NLP）领域上取得极大成功的基于自注意力机制 (Wolf et al., 2020) 的 Transformer 也被提出用于时间序列预测任务上，比较成功的有 Informer(Zhou et al., 2021)、Autoformer(Wu et al., 2021)、Pyraformer(Liu et al., 2022) 和 FEDformer(Zhou et al., 2022) 等。这类都是在 transformer 结构的基础上改进而成。但最近的一项研究 (Zeng et al., 2023) 提出，transformer 这类的模型不能很好地刻画时间方面的信息，在实验中他们发现 transformer 的效果都不如一个非常简单的线性模型 LSTF-Linear。

上述机器学习方法基本上聚焦于只考虑单变量的时间序列预测，而目前多元时间序列预测任务上也提出了许多方法。最近提出来的 STGNN 是一种结合了 GNN 的序列模型，通过多变量序列构图和 Transformer 改进达到了很好的效果，但其同时也具有较高的时空复杂度。此外还有研究基于时序数据的时空不可分性质，设计了 STID 模型 (Shao et al., 2022) 通过使用 id 特征来代替复杂的图模型。

3. Method

3.1. Data preprocessing

对原始数据集，我做了如下处理：

- 将原始数据集进行划分。我将前 70% 的数据用作训练集，后 20% 的数据用作测试集，中间的 10% 数据用作验证集，验证集主要用于模型早停。

[1]Nanjing Unversity.

Time Series Analysis Class Project, Autumn, 2022.

- 对数据做标准归一化:

$$y'_t = \frac{y_t - \mu}{\sigma}.$$

我用从训练集学习出来的均值和方差：μ 和 σ 来对全部数据做标准归一化，使得模型更好学习。

3.2. An improvement: DSTLinear

在分析 Transformer 和 Linear 效果的工作中 (Zeng et al., 2023)，他们提出来一个包含两个线性模型的 $DLinear$，取得了不俗的效果。$DLinear$ 将原始的时间序列分解成趋势项（使用移动平均核来提取）和季节项（原始和趋势项的差值）两个部分，然后用两个线性层分别预测后相加得到最终预测结果。我设计的模型 $DSTLinear$ 就是基于 $DLinear$ 改进的。

我首先沿用了 DLinear 的设计，对原始时间序列趋势-季节项分解。然后考虑到趋势项蕴含着丰富的时空信息，于是选择对趋势项分别提取时间嵌入和空间嵌入。

时间嵌入： 考虑到 TCN(Bai et al., 2018) 的因果卷积和膨胀卷积机制，可以很好地综合不同时间跨度的信息，因此这部分我选择使用 TCN 来提取，即将趋势项输入到 TCN 中来获得时间嵌入。

空间嵌入： 这部分我的实现是一个 $D \times D$ 的线性层（D 为数据的维度），即将趋势项按维度输入该线性模型来得到空间嵌入。

最后将趋势项和上述得到的时间嵌入和空间嵌入加权相加（以类似残差的方式）后送入一层线性层，得到趋势项的预测。

这里为了使得数据的维度一致方便相加，时间嵌入和空间嵌入的大小我都设置成为 $L \times D$。而之所以使用加权的方式相加，是因为我在实验中观察到当预测的未来长度 H 逐渐变长后，模型效果相对于原始的 $DLinear$ 有显著的下降，猜测是过拟合造成，于是采取根据 L、H 的大小动态调整权重的方式来进行模型训练。

假设长度为 L 的 D 维历史数据为 $x \in R^{L \times D}$，模型预测得到的长度为 H 的 D 维未来数据为 $\hat{x} \in R^{H \times D}$，则上述过程可以形式化表示如下:

$$\hat{x}^{H \times D} = W_S \times x_S^{L \times D} + W_T \times (x_L^{L \times D} + k_1 x_{temp}^{L \times D} + k_2 x_{sp}^{L \times D})$$

其中:

- W_S 和 W_T 分别是对季节项和趋势项的线性层
- $x_S^{L \times D}$ 和 $x_T^{L \times D}$ 分别是原始数据 $x^{L \times D}$ 分解后的季节项和趋势项，即 $x^{L \times D} = x_S^{L \times D} + x_T^{L \times D}$
- $x_{temp}^{L \times D}$ 和 $x_{sp}^{L \times D}$ 分别是从趋势项里得到的时间嵌入和空间嵌入
- k_1 和 k_2 是和 L、H 的值有关的权重因子

$DSTLinear$ 总体流程如下:

- 对原始时间序列做趋势-季节项分解
- 季节项直接通过一层线性层得到季节项的预测
- 对趋势项做处理:
  - 对趋势项过一层 TCN，提取不同时间跨度的信息，即获得时间嵌入
  - 对趋势项过一层线性层，提取不同变量间的信息，即获得空间嵌入
  - 将时间嵌入和空间嵌入和原始的趋势项按系数叠加，送入线性层，获得趋势部分的预测
- 将趋势部分和季节部分预测结果相加，得到最终的预测结果

4. Experiments

4.1. Experimental Setup

数据集: 数据集我使用了包括 $weather$、$electricity$、$exchange\_rate$、$national\_illness$、$ETTh1$、$ETTh2$、$ETTm1$、$ETTm2$ 在内的 9 个不同的数据集。

评价指标: 评价指标我使用 MAE 和 MSE。

模型训练方式: 损失函数使用 MSE，$epoch$ 设置为 20，历史数据的长度 L 设置为 336，预测数据长度 H 设置为 {96,192,336,720}。(在 $national\_illness$ 数据集中历史数据的长度 L 设置为 104，预测数据长度 H 设置为 {24,36,48,60}) 在模型训练时采取早停策略（每训练完一个 $epoch$ 就在验证集上评估，若连续 3 次在验证集上效果下降就停止训练），整个实验是放置在 3090Ti 进行的。

4.2. Comparison with LSTF-Linear

在 9 个不同数据集上的结果统计见表1。其中 $DSTLinear$ 的数据是我的实验结果，其余的 $Linear$ 的实验数据摘抄自 (Zeng et al., 2023)。

可以看到，$DSTLinear$ 在数据集 $electricity$、$exchange\_rate$、$weather$ 和 $ETTm1$ 上取得较好

的效果，但是在部分数据集，如 $traffic$、$ETTh2$ 模型性能相比原始的 Linear 有较为明显的下降。

针对这种实验现象，猜测是由如下几个原因造成：

训练集和测试集的数据分布不一致：由于训练数据取的是数据集的前 70% 时间跨度的数据，而测试集取的是数据集的后 20% 时间跨度的数据，有理由相信两者并不是同分布的。同时，在每跑完一个 epoch 的输出日志里可以看到，训练损失和验证损失都会降得较低，但测试损失显著高于前两者。图1可视化地展示了部分数据集的训练、测试分布现象。

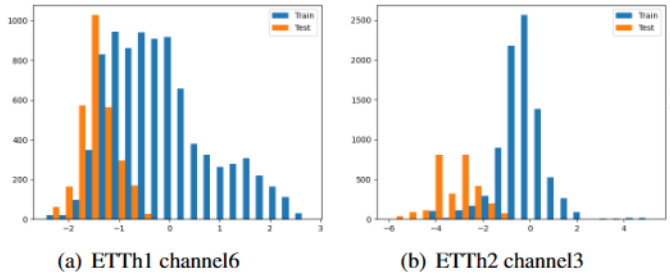

Figure 1. 不同数据集的训练、测试数据分布示意图，图源自 (Zeng et al., 2023)

时空嵌入的维度大小设置不合理：在模型方法设计时，为了使得趋势项及其提取到的时空嵌入可以以类似残差的形式相加，我将时间嵌入和空间嵌入的维度都设置为和原始数据一致的维度了。而不同数据集的维度不同，因此在不同数据集上学得的时空嵌入维度大小也不一样，这也很可能是造成不同数据集上 $DSTLinear$ 的效果有明显不同的原因。$DSTLinear$ 表现得不好的数据集如 $traffic$ 是 862 维，$ETTh$ 数据集是 7 维，很可能对于 $DSTLinear$ 来说，这样的维度不利于表示时间和空间的信息。而 $DSTLinear$ 表现得较好的数据集如 $weather$ 是 21 维，可能适合的时空嵌入维度大小应该在 21 附近。

DSTLinear 的模型结构可能较难学习到空间嵌入：时空嵌入分别是将趋势项经过 TCN 和线性层学习到的。空间嵌入仅根据一个线性层来学习，或许一个线性层不足以来学习到这个复杂的空间嵌入表示。$DSTLinear$ 表现得较好的数据集如 $weather$ 可能各维度之间有比较简单的关联，一个线性层足以学习到；而表现得不好的数据集如 $traffic$ 可能各维度之间的关系比较复杂，一个线性层很难学习到其潜在的关系。

DSTLinear 较容易过拟合：在实验中，我发现如果加深 TCN 和线性层的网络层数，最后的效果反而会变差，这可能说明相对简单的 $Linear$,$DSTLinear$ 很容易就过拟合了。

4.3. Complexity Analysis

$DSTLinear$ 仅由 1 个 TCN 和 3 个线性层（空间嵌入、季节项变换、趋势项变换）组成，因此模型的时空复杂度并不大。$DSTLinear$ 及其他 $LSTF-Linear$(Zeng et al., 2023) 的参数量和推理时间见表2。

5. Conclusion and Future Work

Concusion.
已有实验发现一个简单的线性模型（LTSF-Linear）就可以达到非常好的预测效果。但该模型将多元时间序列看作多个一元时间序列，多个维度共享同一个权重，忽略了变量之间的关联。对此，我在他们提出的 $DLinear$ 模型做出改进，设计了 $DSTLinear$ 模型，将原始时间序列分解为趋势项和季节项，然后对趋势项提取时间和空间嵌入，综合不同维度、不同时间跨度上的信息，以求在较小的时空复杂度里达到更好的预测效果。

Future Work.
$DSTLinear$ 在部分数据集上取得较不错的效果，但是在某些数据集上的效果一般，仍有值得改进的空间。如何设计模型和结构以便更好地学习到时空嵌入，以及如何小心地在欠拟合与过拟合之间找到微妙的平衡，是两个值得继续思考的问题。接下来会继续沿着通过时间嵌入和空间嵌入来综合时空信息的思路来继续改进 $DSTLinear$，使其在多元时间序列任务上表现更好的性能。

Acknowledgments

这是我付出时间最多、收获最大的一门课。说实话，一开始选这课的时候我还是有些犹豫的。因为我只是一个大三的本科生，也没有太多的机器学习基础，对时间序列分析这个领域也完全是一无所知，不得不说跨级上这门课确实有挺大压力。

但还好我坚持下来了。尽管一开始看到课程群里几乎满屏的研究生确实让我倍感压力，但听完叶翰嘉老师的第一堂课之后我便相信我选这门课是一个正确的决定。叶老师的每一次授课都让我学习到很多新内容，课上的每一个方法拓展、论文分享、方法介绍都能让我大开眼界，这也是自从我读大学以来第一次真真切切地意识到我现在在学的这门课是多么的有用，和近年来前沿的内容是有多么的贴近。而课程的作业，感觉也是对我非常友好的，在每一次的作业实践上，都能加深我对这个领域的认识和理解。这学期的所有线下课程，远在鼓楼校区的我也是每一节课都必跑到仙林来上课，来坐到仙 II-110 的第一排座位上。对我来说，这远不仅是一门普通的选修课。

最后的这个大作业，让我这个小白也体验到一丝丝科研的感觉：阅读许多他人的论文，分析他人的实现代码，绞尽脑汁地想各种方法的可行性，不断尝试自己的想法，评估着刚跑出来的数据，以及，还有这篇用 ICML 模板写出来的实验报告。尽管最后的效果可能并不出色，但是这也确确实实给了我很多的收获和很大的成就感。这里面，给了我太多太多从未体验过的感受：对不断补充每一个必需知识的迫切和渴望，对前人精妙想法和设计的钦佩和赞叹，对自我感觉设计的不错的模型跑出来的效果却不如 baseline 的无奈和沮丧，对开始跑模型后还没出来的实验结果的紧张和期待，对模型效果有所好转的兴奋和喜悦。这一切，在别的课上，是完全没有机会体验到的。

尽管这学期因为疫情，线下上课的机会并不是很多，最后也草草以线上课的形式收尾了。但对我来说，这已然足以给我留下非常美好的回忆和体验。最后再次对《时间序列分析》的授课老师和课程助教表示深深的感谢和崇高的敬意！谢谢！

References

Bai, S., Kolter, J. Z., and Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018.

Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., and Dustdar, S. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2022.

Shao, Z., Zhang, Z., Wang, F., Wei, W., and Xu, Y. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information Knowledge Management*, pp. 4454–4458, 2022.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with Auto-Correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. Are transformers effective for time series forecasting? 2023.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pp. 11106–11115. AAAI Press, 2021.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning (ICML 2022)*, 2022.

Table 1. 不同模型在 9 个不同数据集上的多元时序预测任务的性能统计

Methods		DSTLinear		Linear		NLinear		DLinear	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	0.147	0.256	**0.140**	**0.237**	0.141	**0.237**	**0.140**	**0.237**
	192	**0.153**	0.264	**0.153**	0.250	0.154	**0.248**	**0.153**	0.249
	336	0.170	0.280	**0.169**	0.268	0.171	**0.265**	**0.169**	0.267
	720	**0.199**	0.301	0.203	0.301	0.210	**0.297**	0.203	0.301
Exchange	96	0.105	0.225	0.082	0.207	0.089	0.208	**0.081**	**0.203**
	192	0.165	0.298	0.167	0.304	0.180	0.300	**0.157**	**0.293**
	336	0.324	0.435	0.328	0.432	0.331	0.415	**0.305**	**0.414**
	720	0.973	0.744	0.964	0.750	1.033	0.780	**0.643**	**0.601**
Traffic	96	1.314	0.754	**0.410**	0.282	**0.410**	**0.279**	**0.410**	0.282
	192	1.041	0.657	**0.423**	0.287	**0.423**	**0.284**	**0.423**	0.287
	336	0.930	0.614	0.436	0.295	**0.435**	**0.290**	0.436	0.296
	720	0.733	0.486	0.466	0.315	**0.464**	**0.307**	0.466	0.315
Weather	96	**0.167**	**0.226**	0.176	0.236	0.182	0.232	0.176	0.237
	192	**0.211**	**0.266**	0.218	0.276	0.225	0.269	0.220	0.282
	336	**0.256**	0.304	0.262	0.312	0.271	**0.301**	0.265	0.319
	720	**0.320**	0.356	0.326	0.365	0.338	**0.348**	0.323	0.362
ILI	24	1.856	0.942	1.947	0.985	**1.683**	**0.858**	2.215	1.081
	36	2.040	0.990	2.182	1.036	**1.703**	**0.859**	1.963	0.963
	48	1.972	0.977	2.256	1.060	**1.719**	**0.884**	2.130	1.024
	60	2.114	1.017	2.390	1.104	**1.819**	**0.917**	2.368	1.096
ETTh1	96	0.400	0.416	0.375	0.397	**0.374**	**0.394**	0.375	0.399
	192	0.421	0.429	0.418	0.429	0.408	**0.415**	**0.405**	0.416
	336	0.475	0.470	0.479	0.476	**0.429**	**0.427**	0.439	0.443
	720	0.497	0.507	0.624	0.592	**0.440**	**0.453**	0.472	0.490
ETTh2	96	0.664	0.623	0.288	0.352	**0.277**	**0.338**	0.289	0.353
	192	0.999	0.712	0.377	0.413	**0.344**	**0.381**	0.383	0.418
	336	0.937	0.701	0.452	0.461	**0.357**	**0.400**	0.448	0.465
	720	1.421	0.906	0.698	0.595	**0.394**	**0.436**	0.605	0.551
ETTm1	96	0.305	0.352	0.308	0.352	0.306	0.348	**0.299**	**0.343**
	192	**0.334**	**0.363**	0.340	0.369	0.349	0.375	0.335	0.365
	336	0.371	**0.386**	0.376	0.393	0.375	0.388	**0.369**	**0.386**
	720	0.432	0.427	0.440	0.435	0.433	0.422	**0.425**	**0.421**
ETTm2	96	0.218	0.333	0.168	0.262	**0.167**	**0.255**	**0.167**	0.260
	192	0.260	0.341	0.232	0.308	**0.221**	**0.293**	0.224	0.303
	336	0.327	0.386	0.320	0.373	**0.274**	**0.327**	0.281	0.342
	720	0.581	0.531	0.413	0.435	**0.368**	**0.384**	0.397	0.421

1. 评价指标为 MAE 和 MSE

2. **粗体**表示效果最好的，下划线 表示效果第二好的

Table 2. 不同模型的复杂度统计

Methods	Parameter	InferenceTime
DSTL	3859.21k	88.18ms
Linear	69.84k	0.06ms
DLinear	139.68k	0.23ms
NLinear	69.84k	0.09ms

推理时间为运行 10 次取平均值，测试环境为 3090Ti