一种用于多元时间序列预测的结合 Transformer 的扩散模型方法

一、方法名称

一种用于多元时间序列预测的结合 Transformer 的扩散模型方法

二、技术领域

作为支持许多下游应用决策的重要工具，时间序列预测在机器学习的商业和科学领域中扮演了一个重要的角色。这些应用包括但不限于金融定价分析、气象模式预测以及许多其他各种领域。

三、方法背景

时间序列是一种按时间顺序排列的观测值序列，是具有时间上的顺序的一系列数据点，广泛存在于经济、气象、工业、医疗等领域。而时间序列预测则是利用当前已观测到的历史观测值，通过对现有时间序列进行分析学习，观察其中的规律和趋势，建立合适的模型，从而对未来的序列值进行预测。时间序列可能表现出上升或是下降的趋势、在一定周期内出现有规律的波动、或是无规律的变化以及一定的随机性。还可将时间序列划分为平稳时间序列和非平稳时间序列。

其中 ARIMA(自回归积分滑动平均模型) 是较为常用的一种经典时间序列预测模型，包括自回归（AR）和移动平均（MA）两个部分，对于非平稳序列使用差分（I）步骤使序列平稳。首先对时间序列进行分析，判断是否平稳，如果不平稳，则需要使用多次差分使其平稳，然后生成差分数据，并绘制自相关函数（ACF）和偏自相关函数（PACF）图，以此确定 ARMA

模型的参数 p 和 q 的初始值。然后使用已经平稳后的数据进行模型参数训练，利用最大似然估计确定 AR 和 MA 部分的系数。再利用测试验证数据，检查模型的过拟合或欠拟合的情况，最后确认较为拟合后，便可使用该模型预测时间序列的未来值。

Transformer 是一种基于注意力机制的深度学习模型架构，由 Vaswani 等人在 2017 年的论文《Attention Is All You Need》中提出。主要用于自然语言处理（NLP）任务，后来也广泛用于计算机视觉、语音处理等其他领域。其中注意力机制是 Transformer 的核心，使模型在处理序列数据时，对不同位置的数据分配不同的权重，对不同位置的数据注意力不同，从而有效地捕捉长序列中的依赖关系，而其中包括自注意力、交叉注意力和多头注意力机制。Transformer 采用的是编码器-解码器架构，编码器负责将输入序列编码成一个固定长度的向量表示，解码器则根据编码器的输出和之前生成的输出序列来生成下一个输出。相比传统的循环神经网络和卷积神经网络，它在捕捉长距离依赖关系方面表现更出色。并且灵活性高，使用范围广，对其中的组件修改调整衍生出了一系列不同的模型。

四、当前方法的缺点及所要解决的问题

当前的一些技术和方法较少关注时间序列预测中的不确定性。在时间序列预测中，不确定性建模至关重要，因为它直接影响下游应用中预测可靠性的评估能力。这些不确定性因素可能会直接影响到下游应用中预测可靠性的评估能力，从而会显著影响到决策的准确性。在本方法中，基于 Transformer 的表征捕捉能力，提出了一个基于 Transformer 的条件分布学习模型，对于历史序列，通过 Transformer 处理生成一个条件表示，作为先验知识，结合扩散生成模型，达到预测的同时，捕捉到时间序列不确定性的目标。

五、详细阐述

本方法是一个结合了扩散生成过程和一个设计完好的 Transformer 结构的新颖模型，被称为 TMDM(Transformer-Modulated Diffusion Models)。该模型主要由以下两个主要组件组成：

组件一：基于 Transformer 的条件分布学习模型

在本模型中，使用了精心设计的 Transformer 模型，包括 Non-stationary transformer、Autoformer 和 Informer 等。通过它来捕捉历史序列 $x_{0:M}$ 中嵌入的信息，并利用这一信息来对潜在变量 z 进行建模，从而通过该潜在变量进而生成条件表示 $\widehat{y}_{0:M}$，这一表示将用于接下来组件的正向和反向过程。

近年来，在点估计时间序列预测任务的研究上已经取得显著进展，针对不同特性的时间序列产出了不同的专用于不同任务设计的 Transformer 模型，直接采用此类模型生成的条件，比依赖自行设计的条件嵌入更加高效。其次是这些专用 Transformer 表现出了强大的条件均值 $E\left[y_{0:M}|x_{0:M}\right]$ 估计能力，将此估计均值作为条件注入到扩散模型过程中，可以使得扩散模型更加得专注于不确定性的估计上，简化生成过程。使用其他设计引入的条件，增加了生成复杂度。

将给定的 Transformer 结构表示为 $T(\bullet)$，以及历史时间序列 $x_{0:M}$，可以通过 $T\left(x_{0:M}\right)$ 捕捉到条件表示，这个表示将作为引导因子逼近 z 的真实后验分布，该过程可被具体定义为以下方程：

$$q\left(z|T\left(x_{0:N}\right)\right) \sim N\left(\widetilde{\mu}_z\left(T\left(x_{0:N}\right)\right), \widetilde{\sigma}_z\left(T\left(x_{0:N}\right)\right)\right)$$

得到学习好的 z 后，我们可以通过以下式子生成条件表示 $\widehat{y}_{0:M}$：

$$z \sim N(0,1) \text{ and } \widehat{y}_{0:M} \sim N\left(\mu_z(z), \sigma_z\right)$$

我们通过神经网络对三个非线性函数 $\widetilde{\mu}_z$、$\widetilde{\sigma}_z$ 和 μ_z 进行建模，并将协方差矩阵 σ_z 初始化为单位矩阵 $\mathbf{I}$。通过这样的方式，我们定义了一个总结了通过精心设计 Transformer 所捕捉到的信息的潜在变量 z，并将其用于生成了条件表示 $\widehat{y}_{0:M}$，为后续扩散模型的正向和反向过程提供了条件。

组件二：条件扩散时间序列生成模型

通过扩散生成模型的正向过程加噪，以及逆向过程去噪实现生成。

不同于普通的扩散模型假设扩散过程的终点 $y_{0:M}^T$，为标准正态分布 $N(0,1)$，该模型将条件表示 $\widehat{y}_{0:M}$ 概括到了 $p\left(y_{0:M}^T\right)$ 中，更好地体现了条件信息，将扩散过程的终点建模为了以下式子：

$$p\left(y_{0:M}^T|\widehat{y}_{0:M}\right) = N\left(\widehat{y}_{0:M}, I\right)$$

其中 $\widehat{y}_{0:M}$ 包含了 Transformer 捕捉到的条件信息，它可以被看作是基于 $x_{0:M}$ 估计条件均值 $E\left[y_{0:M}|x_{0:M}\right]$ 的先验知识。正向过程，即扩散过程，是对干净数据不断加以噪声扰乱，最终到一个纯噪声分布。对于扩散时间调度 $\{\beta^t\}_{t=1:T} \in (0,1)$，其他时间步的前向扩散过程的条件分布可以表示为以下式子：

$$q\left(y_{0:M}^t|y_{0:M}^{t-1},\widehat{y}_{0:M}\right) \sim N\left(y_{0:M}^t|\sqrt{1-\beta^t}y_{0:M}^{t-1}+\left(1-\sqrt{1-\beta^t}\right)\widehat{y}_{0:M},\beta^t I\right)$$

而在实际过程中，通过式子嵌套计算，化简，我们可以通过以下式子直接从 $y_{0:M}^0$ 采样到任意时间步 t 的 $y_{0:M}^t$：

$$q\left(y_{0:M}^t|y_{0:M}^0,\widehat{y}_{0:M}\right) \sim N\left(y_{0:M}^t|\sqrt{\alpha^t}y_{0:M}^0+\left(1-\sqrt{\alpha^t}\right)\widehat{y}_{0:M},\left(1-\sqrt{\alpha^t}\right)I\right)$$

在这个式子中，我们定义 $\overline{\alpha}^t$ $1-\beta^t$,α^t $\prod_{t=1}^{T}\overline{\alpha}^t$。在从 $t-1$ 到 t 时间步的扩散过程的均值项中，扩散过程可以被概念化为真实数据 $y_{0:M}^0$ 和条件表示 $\widehat{y}_{0:M}$ 之间的插值过程，从真实数据 $y_{0:M}^0$ 出发，逐渐过渡到 $\widehat{y}_{0:M}$。

对于逆向过程，即去噪过程，通过模型学习噪声，并进行逆向去噪，一步一步将噪声数据恢复到干净数据，从而实现数据的生成过程。同样的，我们也将条件表示 $\widehat{y}_{0:M}$ 纳入了逆向过程中，从而利用了历史数据的条件表示，对应的逆向过程式子可以表示为：

$$q\left(y_{0:M}^{t-1}|y_{0:M}^0,y_{0:M}^t,\widehat{y}_{0:M}\right) \sim N\left(y_{0:M}^{t-1}|\gamma_0 y_{0:M}^0+\gamma_1 y_{0:M}^t+\gamma_2\widehat{y}_{0:M},\widetilde{\beta}^t I\right)$$

其中，

$$\gamma_0=\frac{(1-\alpha^t)\bullet\sqrt{\overline{\alpha}^{t-1}}}{1-\overline{\alpha}^t},\gamma_1=\frac{\left(1-\overline{\alpha}^{t-1}\right)\bullet\sqrt{\alpha^t}}{1-\overline{\alpha}^t},$$

$$\gamma_2=1+\frac{\left(\sqrt{\overline{\alpha}^t}-1\right)\bullet\left(\sqrt{\alpha^t}+\sqrt{\overline{\alpha}^{t-1}}\right)}{1-\overline{\alpha}^t}$$

参考图一，为本用于多元时间序列预测的结合 Transformer 的扩散模型方法的流程框架示意图 1，具体实现有如下四个步骤：

1. 对前向过程以及逆向过程的式子进行定义，其中逆向过程包括 t 时间步到 $t-1$ 时间步以及最后 $t=1$ 时的不同处理的定义。使用精心设计

好的基于 Transformer 的模型，包括有 Autoformer 以及 Informer 等模型，作为条件生成模型进行使用，捕捉历史序列的条件信息，以及对于扩散模型噪声预测的模型定义，和对一些损失函数的定义。

2. 采样时间 t，使用刚才采用的 Transformer 条件模型，将历史序列输入进行处理，得到条件预测的结果输出 $\widehat{y}_{0:M}$，再计算 y 和 $\widehat{y}_{0:M}$ 的损失函数，从而对于 Transformer 条件模型进行训练优化。再令 T 时刻 y_T 的均值等于 $\widehat{y}_{0:M}$，即将 $\widehat{y}_{0:M}$ 作为先验均值。

3. 即扩散过程的前向过程，先采样噪声 ε，再通过先前定义的前向过程扩散式子进行扩散计算，其中具体式子定义如下所示：

$$y_{0:M}^{t}=\sqrt{\alpha^{t}}y_{0:M}^{0}+\left(1-\sqrt{\alpha^{t}}\right)\widehat{y}_{0:M}+\sqrt{1-\alpha^{t}}\varepsilon$$

并且再输入进噪声预测模型进行噪声预测，再同先前的条件生成模型的损失函数一同进行损失函数计算噪声 loss，进行训练。

4. 即扩散过程的逆向生成过程。首先在标准正态分布中，随机采样一个噪声 z。再采样时间 t，再计算上述定义过的三个逆向过程式子的三个系数：γ_0，γ_1 和 γ_2，再用训练好的模型进行噪声预测。通过前向过程的式子进行 y_0 重参数化计算，具体式子如下：

$$\mathbf{Y}_{0:M}^{t}=\left(1/\sqrt{\alpha^{t}}\right)\left(y^{t}-\left(1-\sqrt{\alpha^{t}}\right)\widehat{y}_{0:M}-\sqrt{1-\alpha^{t}}\varepsilon_{\theta}\left(y^{t},\widehat{y}_{0:M},t\right)\right)$$

再计算作为后验均值的 $\overline{y}^{t-1}$, 具体计算式子如下：

$$\overline{y}_{0:M}^{t-1}=\gamma_0\bullet\mathbf{Y}_{0:M}^{t}+\gamma_1\bullet y_{0:M}^{t}+\gamma_2\bullet\widehat{y}_{0:M}$$

当 $t\neq 1$ 时，可以得到最后的输出 $y_{0:M}^{t-1}$，即式子为：

$$y_{0:M}^{t-1}=\overline{y}_{0:M}^{t-1}+\sqrt{\widetilde{\beta}^{t}}\bullet z$$

其中，$\widetilde{\beta}^{t}$ 作为后验方差，其计算式子如下：

$$\widetilde{\beta}^{t}=\frac{1-\overline{\alpha}^{t-1}}{1-\overline{\alpha}^{t}}\left(1-\alpha^{t}\right)$$

当 $t=1$ 时，则不用加噪声，即

$$y_{0:M}^{0}=\left(1/\sqrt{\alpha^{1}}\right)\left(y^{1}-\left(1-\sqrt{\alpha^{1}}\right)\widehat{y}_{0:M}-\sqrt{1-\alpha^{1}}\varepsilon_{\theta}\left(y^{1},\widehat{y}_{0:M},1\right)\right)$$

本文采用的目标函数可包括为两个部分，包括扩散模型的部分以及条件生成模型的部分，式子可以表示为如下：

$$\begin{aligned}L_{ELBO}=&\mathbf{E}_{q}\left[-\log p\left(y_{0:M}^{0}|y_{0:M}^{1},\widehat{y}_{0:M}\right)\right]+\mathbf{D}_{KL}\left(q\left(y_{0:M}^{T}|y_{0:M}^{0},\widehat{y}_{0:M}\right)||p\left(y_{0:M}^{T}|\widehat{y}_{0:M}\right)\right)\\&+\sum_{t=2}^{T}\mathbf{D}_{KL}\left(q\left(y_{0:M}^{t-1}|y_{0:M}^{0},y_{0:M}^{t},\widehat{y}_{0:M}\right)||p\left(y_{0:M}^{t-1}|y_{0:M}^{t},\widehat{y}_{0:M}\right)\right)\\&+\mathbf{E}_{q(z|T(x_{0:N}))}\left[-\log p\left(\widehat{y}_{0:M}|z\right)\right]+\mathbf{D}_{KL}\left(q\left(z|T\left(x_{0:N}\right)\right)||p(z)\right)\end{aligned}$$

其中前两行来自扩散模型的损失函数，最后一行源自条件生成函数的损失。

六、技术优点

与现有的多元时间序列预测模型相比，本申请具有以下优点：

1. 通过使用 Transformer 等模型，对历史序列处理，生成条件信息，作为先验知识，并引入到加噪去噪过程中。针对当前模型对于预测过程中对于不确定性忽视的缺点，使用这样的方法进行处理，实现了捕捉不确定性的目标，从而实现了对未来时间序列的高精度分布估计。

2. 本模型在一个统一的贝叶斯框架内，整合了扩散模型和基于 Transformer 的模型。通过采用混合优化的策略，同时对两部分模型进行训练优化。同时作为一个即插即用的模型，可以兼容多个不同的基于 Transformer 模型。

3. 在多个真实世界数据集上，探索了预测区间覆盖概率（PICP）和分位数区间覆盖误差（QICE）作为评估指标，以及其他指标下，表现均很出色，展现了强大的多元时间序列预测的强大能力。