Traditional Method of Time Series Prediction

陈鸿煜

Update date: 2023/6/7

时间序列多步预测指的是根据现有数据计算一个模型去预测未来若干时间步的取值。

时间序列（time series）是一组按照时间发生先后顺序进行排列的数据点序列。通常一组时间序列的时间间隔为一恒定值（如1秒，5分钟，12小时，7天，1年），因此时间序列可以作为离散时间数据进行分析处理。

一、平稳性

平稳性是时序分析的重要概念，时序分析基本上是以**平稳时间序列**为基础的。

平稳性包括弱平稳与强平稳，其中强平稳的证明很困难。一般不需要关注强平稳，常见的时序分析任务都是基于弱平稳做的。

弱平稳需要满足以下三点：

- 均值（即从t=0到当前时间步的均值）为常数
- 方差收敛
- 协方差仅与时间间隔有关，与位置无关。

从统计学的角度讲，平稳性的要求就是对于一个时间序列（分布未知），**这个时间序列的取值一定满足一个确定的分布。**比如我们知道任意一个时间点的取值只能为集合（1, 2, 3）中的某一个数字，取值概率分别为（0.3，0.3，0.4），那么我们认为这个时间序列是平稳的。但是如果在某一个时间点出现了数字4，则从该时刻开始，我们认为时间序列就不是平稳的了。

为什么时序预测需要数据具有平稳性？我们分析时间序列数据是希望能捕捉到数据当中的规律，基于规律，我们可以做预测。因此，平稳可以理解为**这个规律是不随时间变化的**，基于此，我们后面的分析和预测才有意义。否则，规律都是变化的，分析不出结果。

在此思想上，**各种平稳性就是从不同的角度刻画规律**，比如弱平稳以均值、协方差不变来刻画规律的稳定性。

那么可以通过时序预测的经典模型：**ARMIA。**来具体地理解为什么时序预测需要数据弱平稳。

1. ARMIA(差分自回归移动平均)

ARIMA是一种非常流行的时间序列预测统计方法，它是差分自回归移动平均（Auto-Regressive Integrated Moving Averages）的首字母缩写。将**自回归模型**（AR）、**移动平均模型**（MA）和**差分法**结合，我们就得到了差分自回归移动平均模型 ARIMA（p、d、q），其中 d 是需要对数据进行差分的阶数。

(1). AR模型

自回归模型（Auto Regression, AR）描述当前值与历史值之间的关系，用变量自身的历史时间数据对自身进行预测。

一般的一步预测P阶自回归模型：

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \cdots + \alpha_p X(t-p) + u(t)$$

其中 $u(t)$ 代表 t 时刻的随机扰动。自回归模型有很多的限制：

（1）时间序列数据必须具有平稳性

（2）自回归只适用于预测与自身前期相关的现象（时间序列的自相关性）

对于限制（1），如果数据不平稳，那么根据历史数据得出的预测模型（α_i）就没有意义；对于限制（2），如果序列 $X(t)$ 是白噪声，那么**任何两个时点的 $X(t)$ 都不相关，序列中没有任何可以利用的动态规律**，AR模型也就失去了作用。

(2). MA（Moving Average） 模型

MA模型用于处理扰动 $u(t)$，如果 $u(t)$ 是白噪声，代表序列中任意两个值之间无相关性，无法进行序列预测。如果 $u(t)$ 非白噪声，那么通常认为它是一个q阶的移动平均：

$$u(t) = \varepsilon(t) + \beta_1\varepsilon(t) + \cdots + \beta_q\varepsilon(t-q)$$

其中 ε 表示白噪声序列，那么 $u(t)$ 与历史值没有关系，而只依赖于历史**白噪声**的线性组合。我的理解是，对于扰动 $u(t)$，将它理解为依赖于白噪声的时间序列，而非与自身过去值相关。当 $X(t) = u(t)$ 时，可以得到 MA 模型：

$$X(t) = \varepsilon(t) + \beta_1\varepsilon(t) + \cdots + \beta_q\varepsilon(t-q)$$

(5). ARMA 模型

将AR（p）与MA（q）结合，得到一个一般的自回归移动平均模型ARMA（p，q）：

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \cdots + \alpha_p X(t-p) + \varepsilon(t) + \beta_1\varepsilon(t) + \cdots + \beta_q\varepsilon(t-q)$$

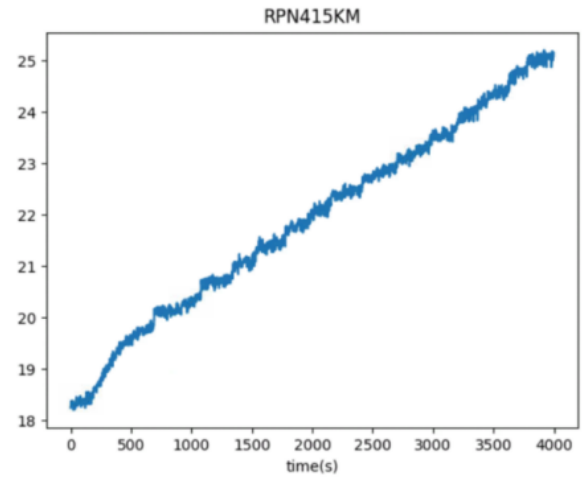
该式表明：

（1）一个随机时间序列可以通过一个自回归移动平均模型来表示，即该序列可以由其自身的过去以及随机扰动项来解释。

（2）如果该序列是平稳的，即它的**行为并不会随着时间的推移而变化**，那么我们就可以通过该序列过去的行为来预测未来。

(6). ARIMA 模型

ARMA 模型只能对平稳序列进行预测，但是对于一些持续增长的时间序列，比如项目里的数据：



（a）

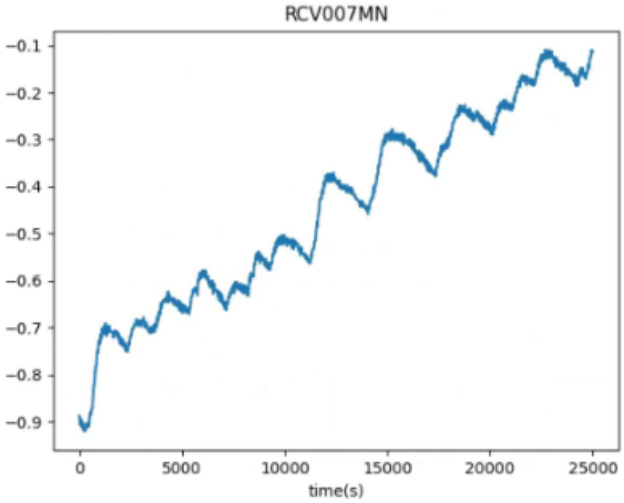
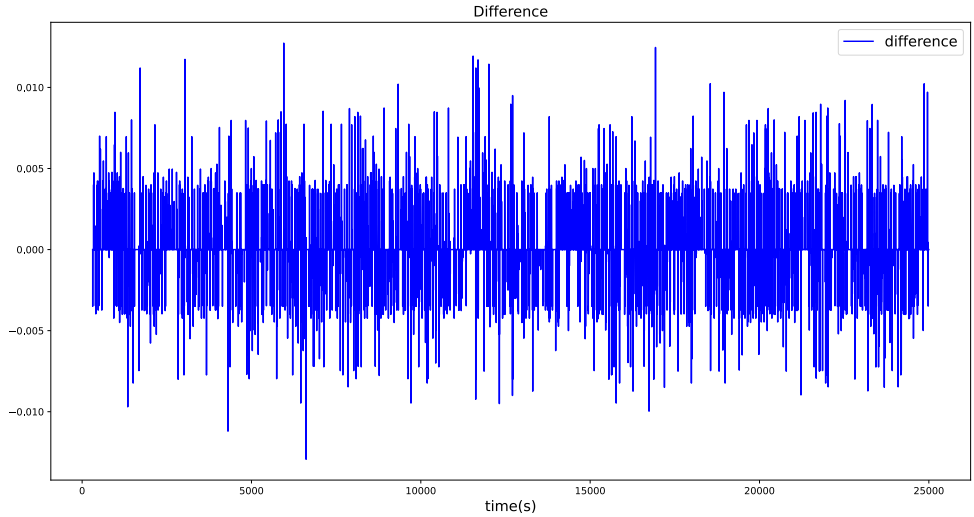


图 6-8 目标变量 RCV007MN 正常工况

这些时间序列非平稳，但是对数据进行差分处理，就可以得到弱平稳序列：



ARIMA模型是通过将非平稳数据差分处理（如果一次差分以后依然非平稳则继续差分），得到平稳序列，再使用AR和MA模型。

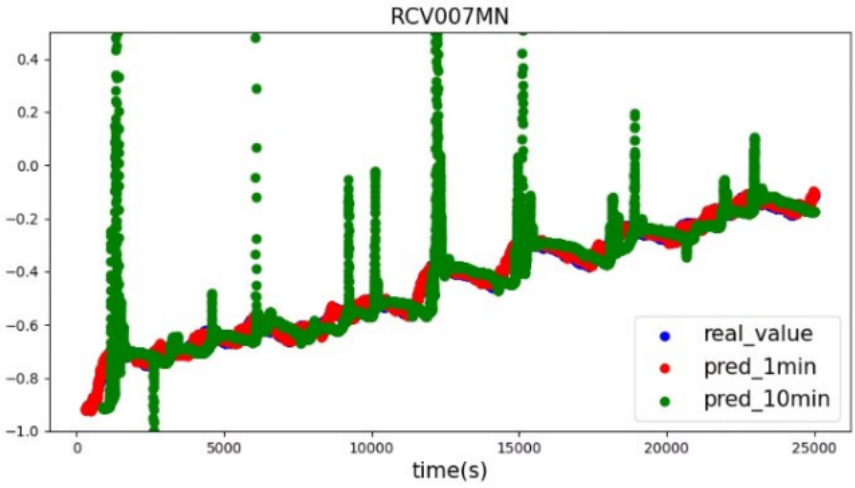
ARIMA模型(p,d,q) 根据原序列是否平稳以及回归中所含部分的不同，ARIMA模型可拆分为3项，分别是**AR(p)模型、I(d)即差分、和MA(q)模型**，因此需要分别确定这三个参数的阶数。一般可使用偏(自)相关图得到合适的p、q阶数，以及使用ADF检验得出合适的差分阶数d。

二、结合项目

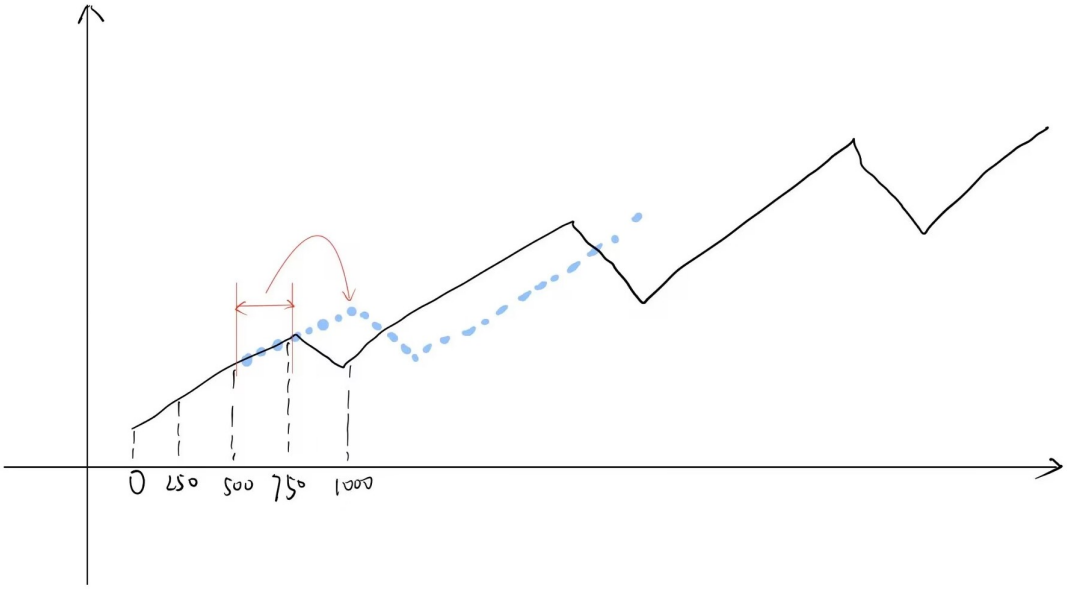
根据给出的仿真数据，是可以使用ARMIA进行时序预测的。但是这只在理想条件下：时间窗口足够大。上图通过25000秒，将近7个小时的时序数据，可以肉眼看出数据的趋势。但是在故障情况下，我们只能用很短的时间序列（项目中使用的是250秒），这么短的时间根本无法进行时序预测，因为无法捕获数据的特征。

这也可以解释为什么近期与**核电站趋势预测/时序预测**相关的论文，都采用了神经网络的方法。神经网络与ARMIA类似，需要长时间的数据来获取数据变化的规律。

在我们的项目中，时间窗口很短：用250秒的过去采样值来预测未来10分钟的值。这就会带来短视的问题，也就是无法获取真实规律。



在这张图中，十分钟的预测虽然看起来只是对应的向后平移。但是结合250秒的时间窗口来理解：



通过500-750内的数据来预测1000秒时的值。之所以看起来是滞后（直接平移），其实是时间窗口太短，捕捉了局部的规律。这也可以解释为什么流行的方法多使用神经网络：先用模拟的数据训练，获取全局规律，再进行预测。

总的来说，当前实现的多入多出自回归方法已经是在当下情况下的比较好的方法。因为只能根据较小量的过去值进行实时预测，无法得到真实的规律。时序预测需要建立在大量数据的经验之下，不符合本项目的应用条件。