时间序列数据异常检测报告

> *15级-李博*
>
> *2017-12-29*

基于统计方法进行时序数据预测的异常检测模型

简述

基于统计的预测，无非是根据收集过去时间的数据，建立一个模型，来计算未来时间的数据，建立的是一种数学或者统计模型，它能表现出已有数据的变化规律，因为大数定理的存在，定义了世间所有的行为都可以通过数字表示，并且存在一定的客观规律。

对于股票市场中存在的量化交易这一概念，即是指以先进的数学模型替代人为的主观判断，利用计算机技术从庞大的历史数据中海选能带来超额收益的多种『大概率』事件以制定策略，极大地减少了投资者情绪波动的影响，避免在市场极度狂热或悲观的情况下作出非理性的投资决策。

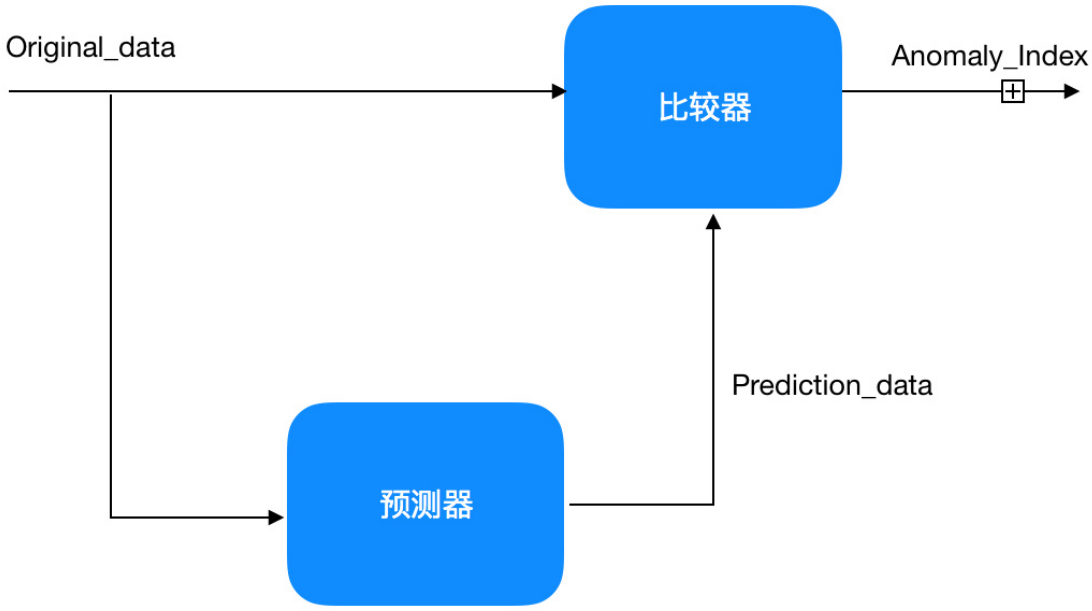
在这里我们从这个方向入手，通过传统股票市场的预测分析工具来进行我们的工业传感器数据分析，首先我简要分析其关联性：

1. 股票数据和传感器数据都具有趋势性，以及一定程度的随机性，趋势性保证了两者的数值会按照一定的规律变化，并且从长时间数据的角度看，其是较为连续的，因此我们可以使用ARIMA，EA等模型进行平滑处理，分析异常点。
2. 股票数据根据金融因素，考虑通货膨胀等原因，可能会存在不断起伏，但是大趋势增长的情况。但是对于传感器数据，大多会在一定范围内震荡（如温度，湿度等数据因为物理因素的原因，不会高出一定范围），所以我们可以在 EA（适用于渐进上升型数据）或者是ARIMA模型（适用于周期波动性数据）中进行决策。
3. 传感器数据可能存在一定的周期，但是股票数据不一定存在明显周期性，这一点也是需要对模型进行修正调整的考虑因素之一。

模型方法介绍

在上一次周报中，我详细介绍了这种方法，现在在这里简单略过。

基于预测的异常检测模型如下图所示，O_{data} 是真实数据，通过预测器得到预测数据，然后 O_{data} 和 P_{data} 分别作为比较器的输入，比较器输出的是真实数据中被判别为异常值的下标 $Index$。



预测器

时间序列分析一般假设我们获得的数据在时域上具有一定的相互依赖关系，通常，如果传感器数值在 t 时刻很高，那么在 $t+1$ 时刻价格也有一定的概率会比较高，而时间序列分析的目的包含以下两个方面：

- 发现这种隐含的依赖关系，并增加我们对此类时间序列的理解；
- 对未观测到的或者尚未发生的时间序列进行预测。

在接下来的分析中，我们认为时间序列 X 由两部分组成，即 $X_t = \hat{X}_t + \epsilon_t$. 其中$\hat{X}_t$ 是有规律的序列而ϵ_t 则无规律的噪声。有规律的 X_t 包含我们想要发现的依赖关系（pattern），而 ϵ_t 我们认为在时间域内不存在相互依赖的关系，即 ϵ_t 和 ϵ_{t+1} 之间是相互独立的。

一个最简单的模型就是我们假设 ϵ_t 是一个随机数，服从一定的概率分布 $f_t(\epsilon)$。

可以发现，我们想要找到 $\hat{X}_t$ 而对 ϵ_t不怎么感兴趣。为了使有规律的 $\hat{X}_t$ 更加明显，我们通常希望能过滤到噪声，而最简单的过滤噪声的方法就是『取平均』。

而问题来了，对于时间序列信号来说，我们该如何取平均呢？这里便引出了我们的基于历史数据平滑曲线的几种方法，也即我们的预测期。

我们使用的预测器主要有以下几种（还在等待其余方法的补充）。

对于时序数据的 F_{t+m} 的预测方法

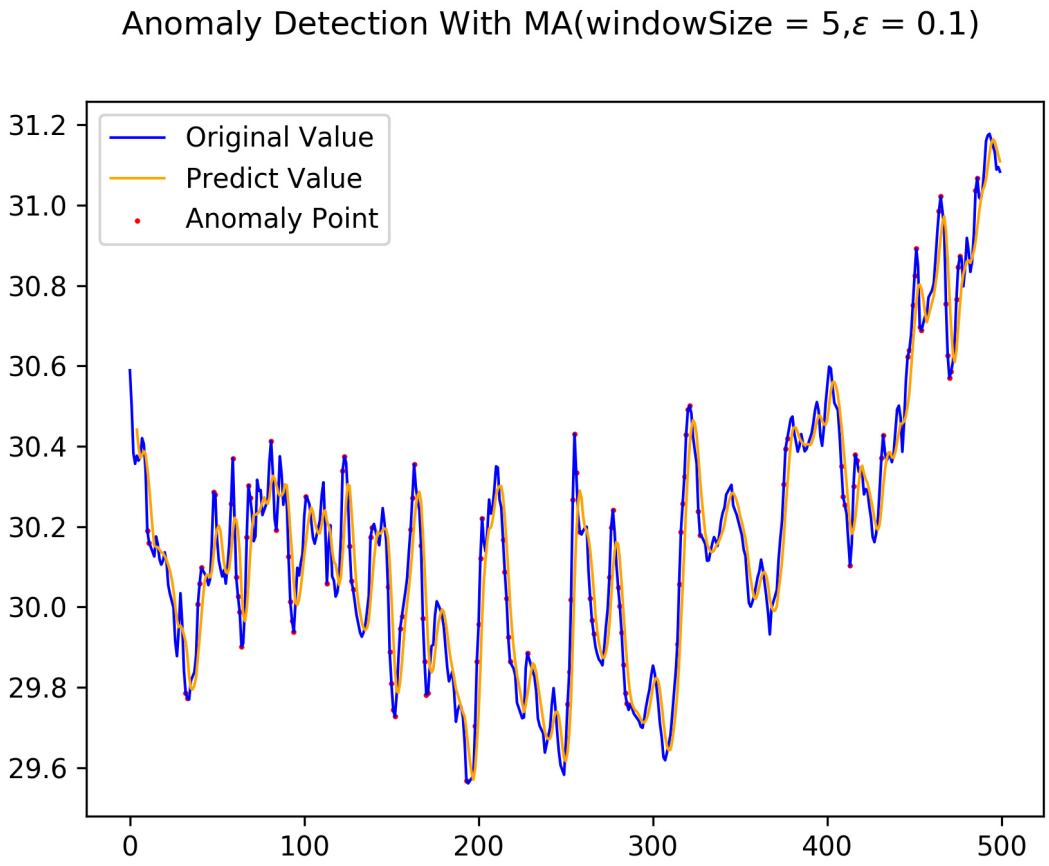
MA(滑动平均模型)

这种方法并不考虑数据的趋势性，单纯根据历史信息来求得当前点的滑动平均数值，参数 T（也即WindowSize） 决定了依赖历史的程度。

我们用 S 表示处理后的序列，那么 S_t 等于 X_{t-T+1} 到 X_t 的平均值，即

$$S_t = \frac{1}{T}\Sigma_{i=t-T+1}^{t} X_i$$

我们使用 $S_t \simeq \hat{X}_t$，下面我们使用滑动平均来预测 windmachine 的第一列的数据，并根据预测值进行异常检测。



EA(指数平均模型)

在这里我们使用二阶指数平均，二阶在一阶的基础上增加了对于趋势的考量，更符合我们的数据要求，而三阶需要依赖于周期性，这点和上述的 ARIMA 模型所依赖的周期性也是相同的，下一步我们也需要加入周期性的考量。

我们可以看到，虽然指数平均在产生新的数列的时候考虑了所有的历史数据，但是仅仅考虑其静态值，即没有考虑时间序列当前的变化趋势。如果当前的股票处于*上升趋势*，那么当我们对明天的股票进行预测的时候，好的预测值不仅仅是对历史数据进行『平均』，而且要考虑到当前数据变化的*上升趋势*。同时考虑历史平均和变化趋势，这边是**二阶指数平均**。 我们先给出二阶指数平均的两种方法，如下

1. $initialize\ S_1 = X_1\ b_1 = X_1 - X_0\ for\ t > 1$
   $S_t = \alpha X_t + (1-\alpha)(S_t - 1 + b_t - 1)\ b_t = \beta(S_t - S_t - 1) + (1-\beta)b_t - 1$
   $end\ for$

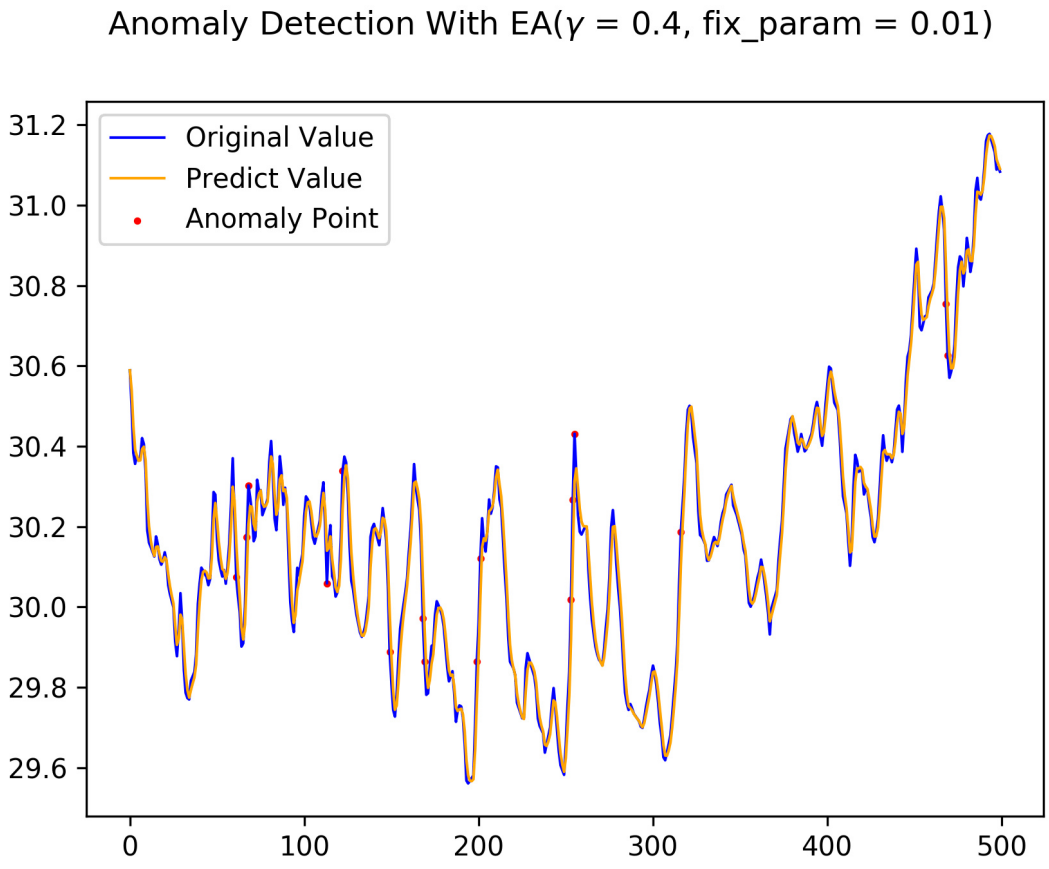
   如果我们对 X_{t+m} 之后的数值进行预测，那么我们的预测值为
   $\hat{X}_{t+m} = S_t + mb_t$

2. $S'_0 = X_0\ S''_0 = X_0\ for\ t \geq 1$
   $S'_t = \alpha X_t + (1-\alpha)S'_{t-1} \quad S''_t = \alpha S'_t + (1-\alpha)S''_{t-1} \quad S_t = 2S'_t - S''_t$
   end for

   在方法二中，只有一个参数 α 。其中S'_t为最基本的指数平均得到的结果，而 $S'_t - S''_t$ 为变化的趋势.

   如果我们对 X_{t+m} 之后的数值进行预测，那么我们的预测值为
   $\hat{X}_{t+m} = S_t + (m\frac{\alpha}{1-\alpha})(S'_t - S''_t)$

我们利用二阶指数平均对于数据进行处理及异常检测的结果如下：



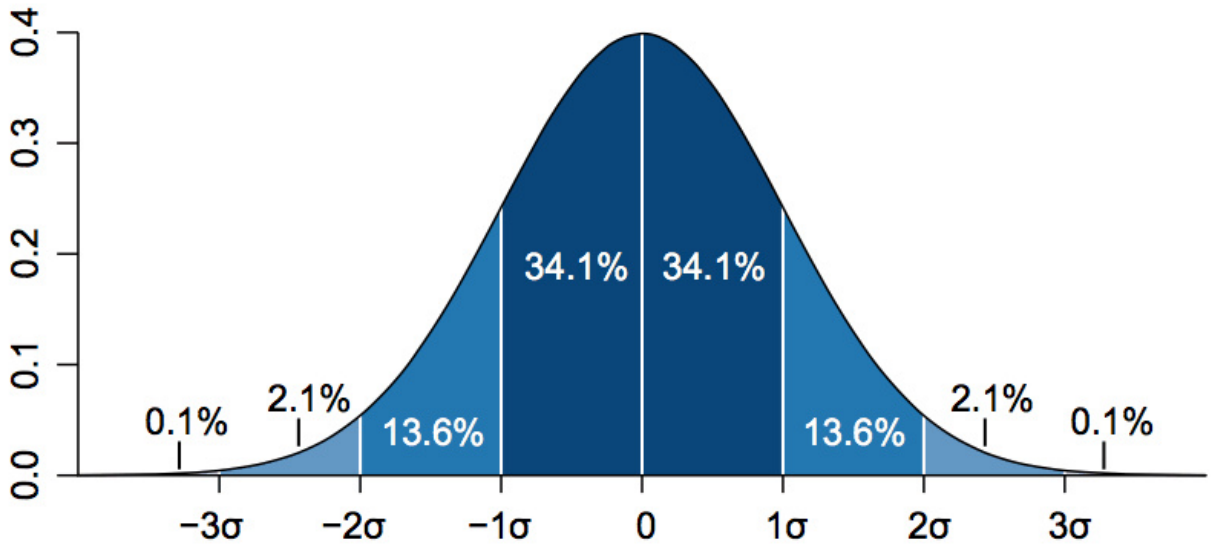
用于修正的方法

布林通道

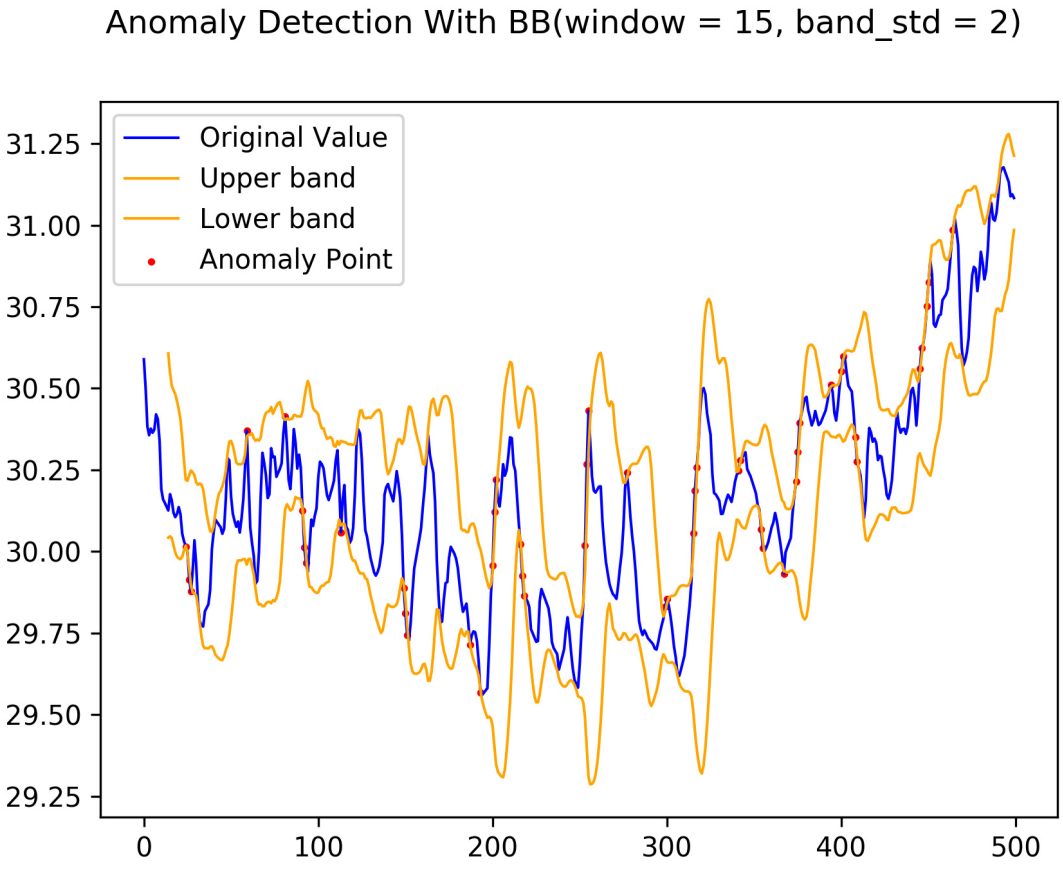
布林通道是一个在股票市场中经常使用的概念，它在应用上结合了移动平均和标准差的概念，其基本的型态是由三条轨道线组成的带状通道（中轨和上、下轨各一条）。上下轨分别由平均值加减二倍标准差（$\pm 2\sigma$）得到，**中轨**为股价的平均成本，**上轨**和**下轨**可分别视为股价的压力线和支撑线。

> *下图中红线为上轨，绿线为下轨，蓝线为滑动平均值。*

由大数定律以及高斯分布模型，我们可以获知在上下二倍标准差基本涵盖了数据变化的95%左右的情况，如果超出这个分布，那么极有可能是异常点，因此在我们的模型中我们结合布林通道进行了异常点的修正。



下面是我们使用布林通道进行修正测试：



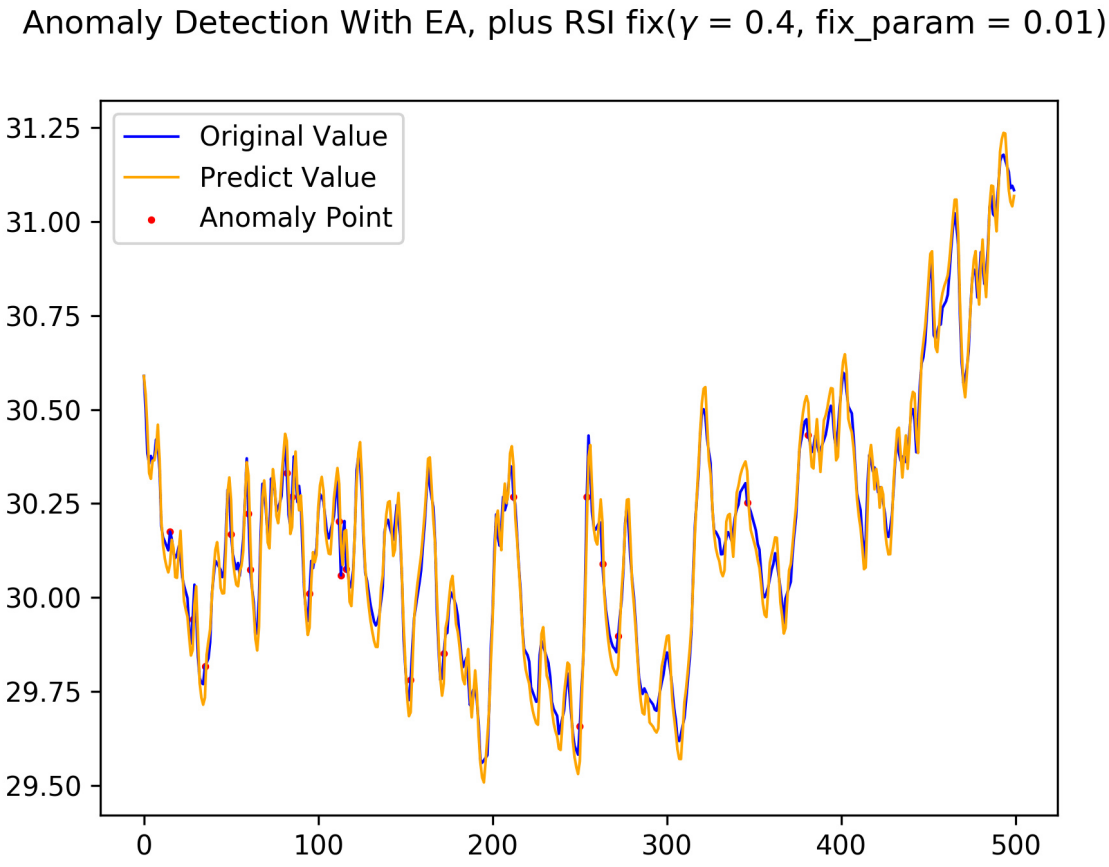
RSI指数

RSI 指数在证券市场中适用于衡量一段周期内增长和下降强弱对比的一个指标，其计算公式如下：

1. RSI = 100×RS / (1+RS)
2. RS = X天的平均上涨点数 / X天的平均下跌点数

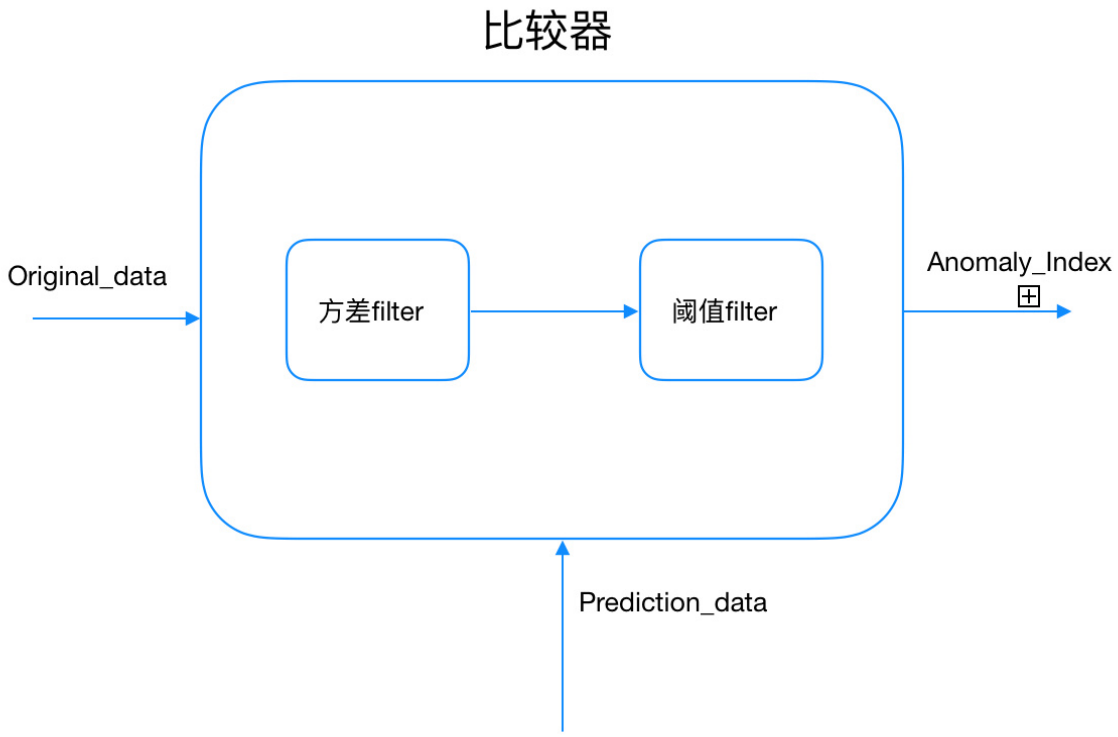
通俗的讲，RSI 指数过高代表上涨明显（超买现象），意味着有可能会存在潜在的下跌，RSI指数过低则反之。

下面是我们使用修正之后的模型结合EA方法进行异常检测的情况：



比较器

预测器预测出当前时刻传感器的预测值后，还需要与真实值比较来判断当前时刻数据是否异常。一般的比较器都是通过阈值法，比如实际值超过预测值的一定比例就认为该点出现异常，进行报警。这种方式错误率比较大。在传感器数值模型的报警检测中没有使用这种方式，而是使用了两个串联的 Filter，只有当两个 Fliter 都认为该点异常时，才进行报警，下面简单介绍一下两个 Filter 的实现。



离散度Filter

根据预测误差曲线离散程度过滤出可能的异常点。一个序列的方差表示该序列离散的程度，方差越大，表明该序列波动越大。如果一个预测误差序列方差比较大，那么我们认为预测误差的报警阈值相对大一些才比较合理。离散度 filter 利用了这一特性，我们对于这个真实数据点的前 N 个点求方差，下一步阈值 filter 的输入为方差 σ.

阈值Filter

根据误差绝对值是否超过某个阈值过滤出可能的异常点。利用离散度 Filter 进行过滤时，报警阈值随着误差序列波动程度变大而变大，但是在输入数据比较小时，误差序列方差比较小，报警阈值也很小，容易出现误报。所以设计了根据方差 σ 进行过滤的阈值 filter。阈值 filter 设计了一个分段阈值函数 $y = f(x)$，对于实际值 x 和预测值 p ，只有当 $|x - p| > f(x)$ 时报警。实际使用中，可以根据数据寻找一个对数函数替换分段阈值函数，更易于参数调优。

总结

创新性

目前少见有人使用预测器 + 判别器的方式结合金融统计方法进行异常检测，这种方式本身比较新颖。主要优势有如下几点：

1. 相对于我们之前采用的机器学习的方法，这种方法即使在大数据的情况下也能够顺利完成，因为根据选取方法的不同，基本都是在 $O(N)$ 级别的方法。
2. 可以实现在线实时预测，可以将成套的算法写入传感器芯片内作为硬件级别的预测，芯片内只需要存储规定 N 日内的数据即可。

不足

1. 在预测器中其实我们也应该考虑ARIMA或者是三阶指数平滑（即Holt-winters）模型，但是对于传感器数据的先验信息我们并不了解，导致我们不能够手动的设置其周期值，对于周期性不明确的情况，我们暂时使用 fix 方法作为补偿。
2. 下一步我计划使用RNN作为预测器，使用 LSTM 方法，直接对数据进行预测，但是这种方法可能会比现有的方法更耗时，但是有可能会有更高的准确性，而且目前我已经在尝试使用 LSTM 提取数据的周期性，效果还不错。

困难点分析

1. 对数据的先验知识不够，这近似于一个非监督学习，如果要对效率有更高的要求可能需要更多的标记。
2. 对未来这个系统的使用方法，使用对象和项目雏形不是很了解。
3. 目前尚未完成对于数据的周期性提取的高效算法（LSTM相对来说比较耗时，对于大数据可能不太适用）。
4. 目前尚未完成对于模式异常的识别。

研究方向

1. 正在学习使用 RNN 进行序列的预测，到时候会根据正确性以及效率来和现有的方法进行选择以及比较。
2. 正在考虑是否需要将大段的点异常归类为段异常，从而进行模式异常的识别。
3. 在目前我们的实验中，主要使用了两种预测方法+两种修正方法，但实际上时序数据还有许多可以增加的模型和修正方法（我们倾向于使用多修正方法进行参数投票，最终用拟合的方式找出一个最适合我们训练数据的函数），这些方法的最终目的就是能够更加根据历史信息平滑的预测变化量，从而为我们的异常检测提出建议的值。