应用lstm算法进行cpu剩余量预测

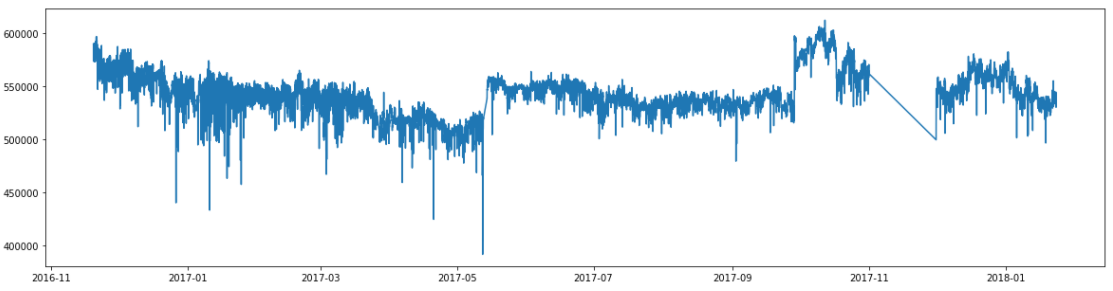
一 数据生成与处理：

1 数据生成

使用数据频率：分钟级数据，即通过每10分钟采集一次生成的数据

cpu剩余量：通过cpu总量与使用量的差值得到cpu剩余量

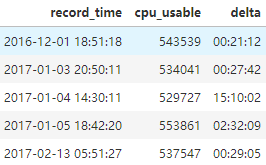
cpu剩余量原始图形如下



2 缺失数据处理

可以看到，数据中存在一些缺失点，为保证每条输入/输出数据的连续性，需要对确实数据进行处理，处理过程如下：

a.找到数据缺失点：计算本条数据与上条数据生成时间的差值t，挑选出t大于15分钟时对应的数据（因原始数据是每个10分钟左右采集一次数据），部分样本点如下（delta即为t）：



发现共有114处存在数据缺失。

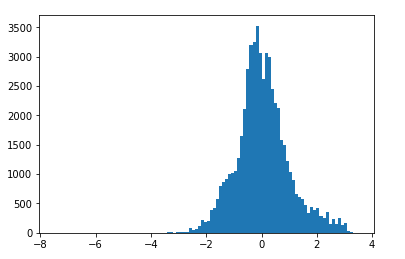
b样本缺失点114处，由于缺失时间段长短不一（20分钟到几十天不等），而且数据变化通常为突变，因此很难通过拟合法等通用方法进行确实数据填补。鉴于总数据共68000余条，舍弃缺失处的一段数据对数据变化量不大，因此对缺失数据进行如下处理：在生成神经网络的输入输出时，若某次输入输出中包含某个缺失数据，则舍弃该输入输出。

3 划分训练集和预测集

从cpu剩余量原始数据中可以看出，数据在2017/11-2017/12月间数据缺失严重，且2017/12后数据量占总数据量约15%，故将2017/12/01之前数据用作训练集，之后数据用作测试集。

4 数据标准化

神经网络系统要求对输入输出数据进行标准化处理。为避免未来函数影响，在训练集上，进行如下标准化处理：（训练数据-训练集数据均值）/训练集数据标准差；在测试集上，进行如下标准化处理：（测试集数据-训练集数据均值）/训练集数据标准差。最终使全部数据得以标准化。最终数据分布如下图：



95%以上数据分布在[-2,2]区间，且数据整体较符合正态分布。

5 输入输出数据划分

目前用前24小时的数据预测十分钟后的数据。即输入为144个数据，输出为1个数据。在原始数据上逐条生成输入输出数据。训练集共46210条，预测集共6333条。

二 模型构建

基础模型为一层lstm层加上一层dense层。

模型中对预测结果影响较为显著的两个参数为lstm层输出维度和激活函数。

lstm层输出维度即lstm层所包含细胞个数，细胞个数越多，内部参数越多。本研究中输出维度取值为16，32，64.

激活函数为每个lstm细胞内处在各个门中影响lstm内部参数优化的一个可变参数。本研究中激活函数分别选取 tanh，2\*sigmoid 两个激活函数。

在此基础上可以做出的改变是增加神经网络层数。鉴于训练数据有限，增加1层lstm层会导致模型内部参数数量接近甚至超过训练数据总量，故本研究增加的仅是dense层的数量，即 1层lstm+1层dense 和1层lstm+1层dense。

根据上面三方面，共生成3×2×2=12个模型。

三 预测结果与真实结果对比

18个模型MSE对比如下(由于每次模型内参数初始化数值不同，每次跑模型会有稍许差异)：

1层lstm+1层dense：

|  |  |  |
| --- | --- | --- |
| 输出维度(纵)\激活函数(横) | tanh | 2\*sigmoid |
| 16 | 0.0343 | 0.0372 |
| 32 | 0.0251 | 0.0275 |
| 64 | 0.0231 | 0.0251 |

1层lstm+2层dense：

|  |  |  |
| --- | --- | --- |
| 输出维度(纵)\激活函数(横) | tanh | 2\*sigmoid |
| 16 | 0.0362 | 0.0412 |
| 32 | 0.0313 | 0.0345 |
| 64 | 0.0286 | 0.0300 |

由结果可知，在其他条件不变的情况下，输出维度越大，mse越小，这与初始的判断是相符的：输出维度越大，lstm内部越复杂，参数越多，拟合的越好。

在本研究问题中，激活函数tanh比2\*sigmoid要好。

单层dense层mse要相对小于两层mse。这一点与初始判断不符，因为dense层增加同样使模型参数变多，但是模型mse反而变大了。原因目前尚不明确，有待分析。

综合上面结果，最终模型确定为输出维度为64，激活函数为tanh，结构为1层lstm+1层dense’层的神经网络模型。

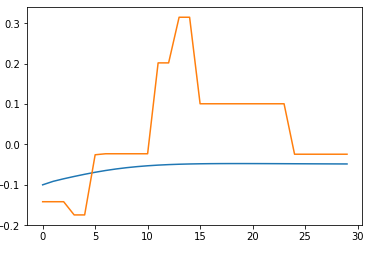
四 利用最终模型进行连续预测

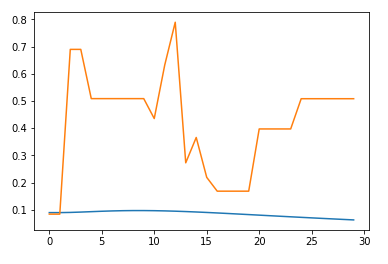
上述过程是利用过去24小时（144个数据）预测未来10分钟（1个数据）.为了达到连续预测，即利用过去24小时（144个数据）预测未来100分钟（10个数据），采取利用上次预测数据进行预测的方法，具体如下：

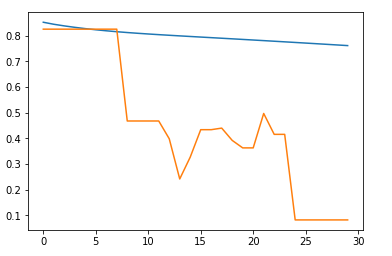
选取连续144个数据构成数据段array\_a，将array\_a作为输入数据输入模型，得到预测结果b。将b并入array\_a最后一位，此时a有145个数据，然后取array\_a的第2至145这144个数据形成array\_b，再输入模型进行预测得到新的预测数据。依此类推，可得到未来一段时间的多个预测数据。

通过多次测试，发现进行连续预测的效果并不好，除预测值序列与实际值序列的第一个值较接近外，预测值序列通常为一个连续渐变过程，实际值序列为非连续突变过程。

部分结果（预测未来300分钟）如下（蓝线：预测值；黄线：实际值）：







五 总结

在10分钟级别神经网络模型对于cpu未来一段时间变化的预测是不成功的，推测原因为在10分钟级别上cpu变化是非连续的和突变的，规律性差。目前所用模型较简单，但是已经能够反映出基本的情况。就目前结果以及原始数据的变化规律性来看，我认为即使将模型变复杂以及调整更多参数，在分钟数据上很大概率得不到积极的结果。日线级别cpu变化规律强一些，但是日线数据太少，无法使用神经网络模型预测。