风速预测：数据科学导引期末大作业报告

赵天洋 1500017799
张宏毅 1500017736
郭嘉明 1500017789

January 11, 2018

1 问题背景

上世纪 70 年代爆发的世界性能源危机，以及近年来气候环境的变化，促使许多国家更加重视风能等新能源的研究、开发和利用。风能作为一种清洁能源，是目前为止可再生能源中利用最为广泛，技术最为成熟，商业化生产及规模化开发方面最具前景的能源之一。但同时，自然界的风具有非常强的随机性和间歇性，这使得风机功率也相应地有非常强的波动性及不可控性，从而使电网调峰、电压及无功控制非常困难，影响了风电在电网中的大规模接入。因而，对风速及风机功率进行评估和预测是评定大型风电项目是否可行的重要工作内容，对一个地区进行准确的风资源计算可为风能参数的适当选择以及系统状态的设计奠定基础，从而提高的可靠性及可控性，对机组功率进行合适的优化，达到风机出力的预期效果，创造最佳的经济效益和社会效益。

在本项目中，我们拟对 2016 年四月至七月的风速进行预测。数据集由三部分组成，第一部分是原始历史数据，即当前日期之前各天的风机风速及相应功率，每天中风速和功率测量的时间点间隔不规律；第二部分是天气预报数据，包括风速、温度、气压等物理数据的预测值、预测的时间点为每 15 分钟一次；第三部分为风机参数数据，包括各风机的满载功率与满载风速等相关参数。利用这些数据，我们拟采用能够提取出时序抽象特征的模型，对这些日期的风速，按照每 15 分钟的间隔进行对应的预测与评估。

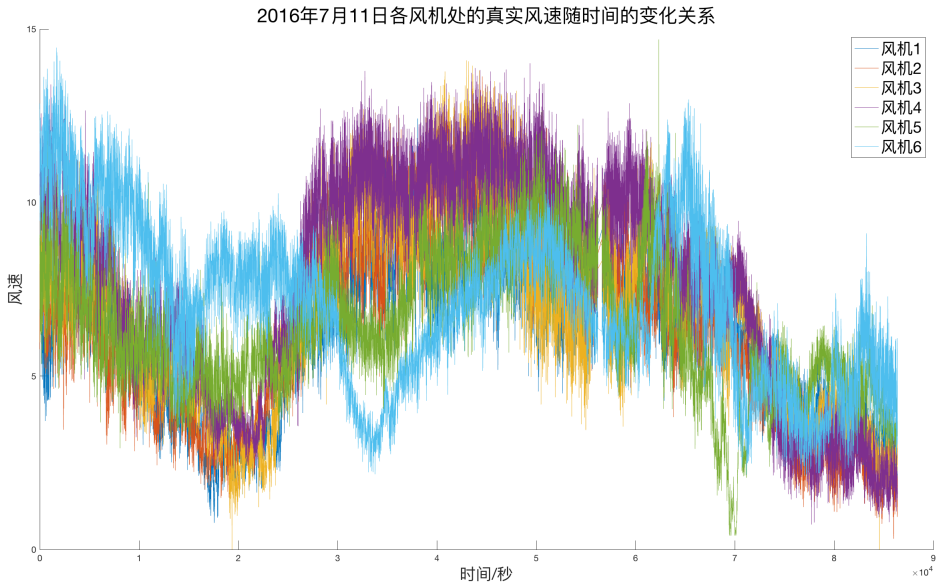


图 1: 一天内各风机处的风速变化趋势

2 数据初步探索

为较全面地了解数据，我们对数据进行了可视化过程。首先，我们任取数据集中的一天，分析各个风机处的风速随时间的变化关系，结果如图 1 所示。

从中可以看出，六个风机处的风速随时间有连续变化的曲线关系。在夜里 23-24 点和清晨 5-6 点，风速达到全天的两个最小值；整个白天中风速均较大，在 11-12 点处风速达到最大值。全天风速范围大致落在 0 到 15 之间，在某些时间处风速测量是间断的。注意到，第六台风机的风速趋势与另外五台区别较大，这可能是由其特别的地理因素导致，但这并不妨碍这些风机处的风速都服从相似的函数关系，且大体呈现出时间上的周期性，为了能够捕捉到这一特别的时序特征，在模型构建中我们考虑使用 LSTM 进行预测，具体方法将在第四部分中叙述。

同时我们还画出了同一天内功率随风速的变化关系，如图 2 所示。从中可见，在风机工作正常的前提下，功率和风速呈正相关。在理想状态下，假设气流所具有的全部动能均转化为风能，则根据功能原理，可以估计出当风速为 v 时，单位时间内垂直通过截面积 S

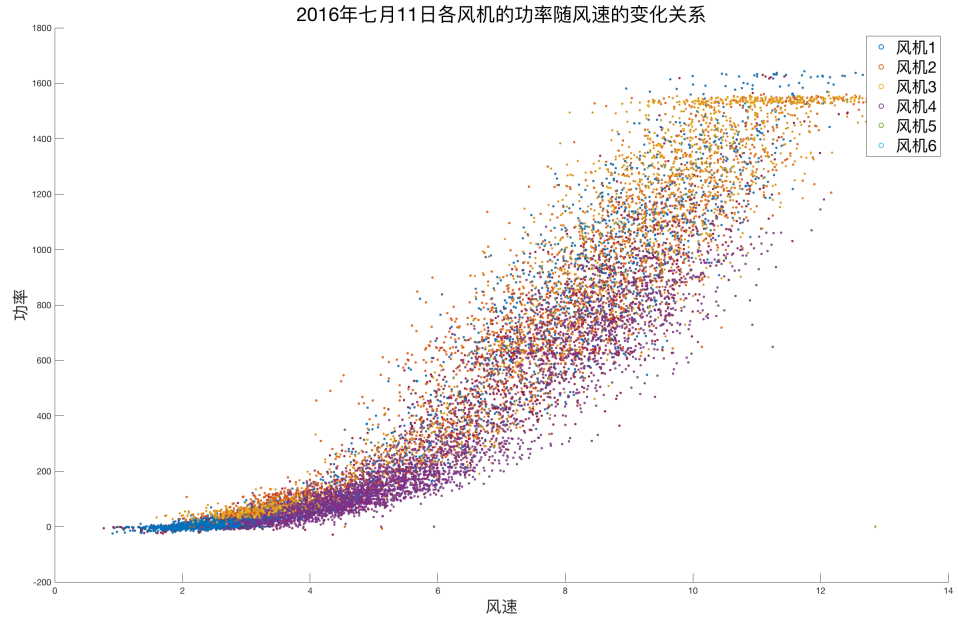


图 2: 风机功率与风速的关系图

的风能为

$$P = \frac{\mathrm{d}W}{\mathrm{d}t} = \frac{1}{2}\frac{\mathrm{d}m}{\mathrm{d}t}v^2 = \frac{1}{2}\frac{\rho S v \mathrm{d}t}{\mathrm{d}t}v^2 = \frac{1}{2}\rho S v^3,$$

其中 ρ 为气流密度，与风机的地理位置有关（包括经纬度、实时气温、实时气压、实时湿度等相关物理因素）。但实际上，气流动能不可能完全转化为电能，Betz 定律给出最终的有效能量约为上述结果的 59.3%。由此可以看出，功率与风速的三次方成正比关系，这一理论结果与实际数据符合得也相当好。

对于数据中天气预报给出的气压和气温，为简单起见，我们以某一日的数据为例，画出了风速和气压、气温间的三维关系，如图 3 所示（其中 x 轴为气压，y 轴为风速，z 轴为温度）。从图中我们并不能发现三者之间明显的相关关系。进一步地，相关研究表明，风速预测的误差中大约有 80% 来源于数值天气预报的不准确性，考虑到这些因素，我们在模型训练中选择不使用天气预报中的气压和气温这两个变量，转而着重研究风速随时间的长期变化情况。若能获得实时的气象数据，则我们相信将这些数据包括在模型中，可以使得模型更加可靠，实现更精确的在线的风速预测。

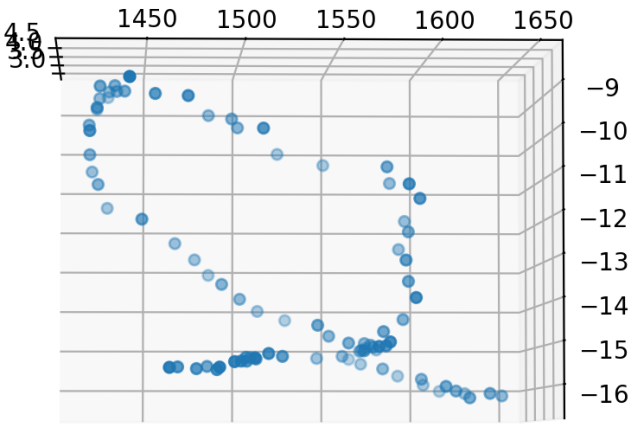


图 3: 风速与预报的气温气压的关系

3 数据预处理

我们需要针对 LSTM 模型构建合适的数据集。假定时序特征主要隐含在当前时间点前的 m 个数据点中，则每个输入的数据点 $\boldsymbol{X}_k$ 为一个矩阵

$$\boldsymbol{X}_k = (\boldsymbol{x}_{k-m}, \boldsymbol{x}_{k-m+1}, \cdots, \boldsymbol{x}_{k-1}) \in \mathbb{R}^{6\times m},$$

其中向量

$$\boldsymbol{x}_{k-j} = (v_{1,k-j}, v_{2,k-j}, v_{3,k-j}, v_{4,k-j}, v_{5,k-j}, v_{6,k-j})^T,$$

表示各个风机在第 $k-j$ $(1 \le j \le m)$ 时刻的风速数据。对应的标签为向量 $\boldsymbol{y}_k$，即六台风机在第 k 时刻的真实风速。具体而言，对于 15 分钟间隔的预测任务，$\boldsymbol{x}_k$ 即表示第 k 个 15 分钟，即第 $900k$ 秒时的风速数据。对于天气预报中各风机处当前时刻的温度、气压等数据，如前所述，我们选择不使用这些气象数据，以避免由于天气预报的误差造成误差放大。

在数据预处理过程中，我们需要考虑应采用数据中的历史真实数据（HisRawData）还是天气预报数据（PredData）来进行模型的训练。一方面，天气预报的风速未必能准确表征真实风速，我们无法得知天气预报是如何进行的，预报中已隐含有一些未知的误差，使用该数据可能会造成误差的累积；另一方面，历史的真实数据又存在数据质量参差不齐、部分观测点间的时间间隔过大（如间隔 30000 秒）、部分月份（如四月、七月等）数据严重

缺失等问题。这里的处理方式我们决定与前一节中一样，选择采用真实风速的数据，并对其插值后求得 $\boldsymbol{X}_k$ 和 y_k 的参考值。针对这些数据，在数据质量不齐，测量点间距较大的前提下，我们考虑使用分段线性插值，即对于时间点 $t_0, t_1, \cdots, t_k$，采用分段线性函数

$$f(t) = \frac{x_i - x_{i-1}}{t_i - t_{i-1}} t + \frac{x_{i-1} t_i - x_i t_{i-1}}{t_i - t_{i-1}}, \quad t_{i-1} \le t \le t_i,$$

来作为不等间隔数据间的填充和近似。当然也可以采用其他的插值方法，例如三次样条插值、B 样条插值等，但考虑到有些观测点之间的时间间隔过大，导致这些插值的结果会有非常大的方差，所以对这些方法，应考虑光滑样条，在样条的基础上对函数二阶导数的积分添加一项正则化项，即

$$\hat{f} = \arg\min_f L(f) = \arg\min_f \sum_{i=1}^{N} [x_i - f(t_i)]^2 + \lambda \int [f''(t)]^2 dt.$$

以控制插值结果的方差，一定程度上保证结果的优良性。

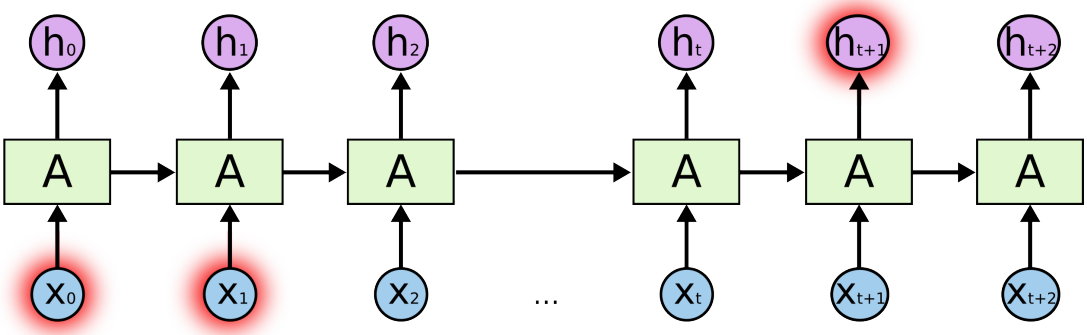
我们尝试先对每一天的数据进行插值，之后再将它们并起来。但注意到每天的数据并不是从 0 秒开始（一般测量起点为第 3 秒至第 7 秒），我们对 0 秒处结果的估计缺少了可依赖的数据项，同时在数据集的构造中，又必须取到第 0 秒的真值作为一个数据点。这启发我们把四个月的所有数据合并到一起进行分段插值，此时每天第 0 秒的数据可以由前一天最后一个测量点和今天第一个测量点进行估计，增加了插值数据的可靠性。

经插值，我们得到了 $T = \{\boldsymbol{x}_k\}_{k=0}^{N-1}$ 这样一个时间序列，其长度 $N = 11616$。为保证训练 LSTM 时梯度能有效地反向传播，考虑取每个输入数据点 $\boldsymbol{X}_k$ 的列数 $m = 60$，则所有可用数据为 $S = \{(\boldsymbol{X}_k, \boldsymbol{y}_k)\}_{k=60}^{N-1}$，共包含 11556 个数据点。从中随机划分训练集和测试集，取训练集大小为 10000，测试集大小为 1556，二者的比值约为 $6:1$，进行后续的模型训练与评估。

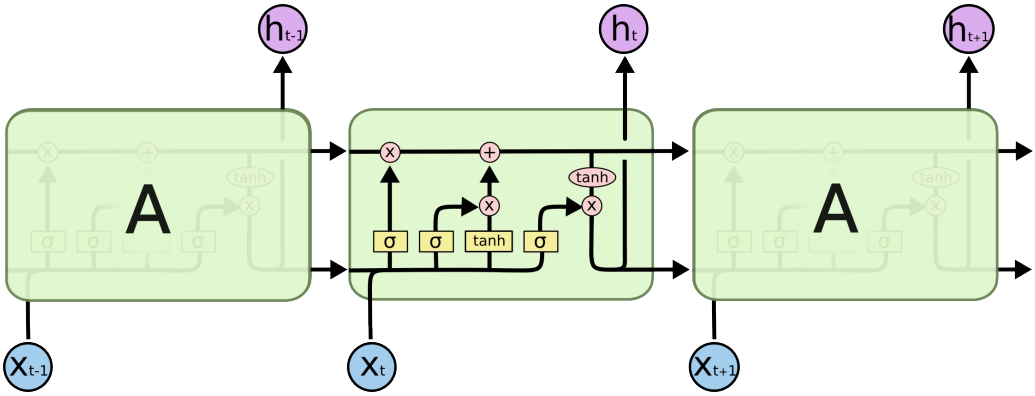
4 模型构建方法

4.1 时序网络模型 LSTM 简介

对于具有内在时序关系的预测任务，循环神经网络（RNN）是一种非常合适的基于人工神经网络的模型结构。与前馈神经网络不同，RNN 在网络中引入了循环，展开结构如



(a) RNN 的典型结构



(b) LSTM 的基本结构

图 4: RNN 与 LSTM 的结构示意图

图 4(a) 所示，其中 A 为隐藏单元（hidden unit）。每个隐藏单元中均有隐藏状态（hidden state），用于接受当前步的数据输入，前一步的隐藏状态输入，并产生该步的输出。由于每个隐藏单元都接受了前面所有时间步的隐藏状态信息，因而可以预期这样的结构能够在序列关系上有出色的表现。

而在风速预测的任务中，核心的困难在于输入数据与实际有效的风速规律之间存在巨大的维数差异，例如对于具有日周期变化规律的风速，如果采用 15 秒为时间间隔进行风速采样，则平均出现一次周期需要 5760 个数据点。但实践表明，若令 RNN 每次预测一个样本点，则模型往往很难在高时间分辨率的任务上有出色的表现，例如在图 4(a) 中。输出 h_{t+1} 可能和它所依赖的输入 x_0 和 x_1（如红色圈所示）之间有非常密集的时间点，那么当

t 充分大时，RNN 的模型效果就会变得不尽人意。鉴于此，我们使用一种特殊的 RNN 结构，即 LSTM（Long-Short Term Memory）来发掘长时程的时间依赖关系。LSTM 采用了一种特别的隐藏单元结构，如图 4(b) 所示。

在 LSTM 单元中，除了 RNN 中已有的隐藏状态 h_t 外，还引入了单元状态（cell state）以表征长时记忆，记为 C_t。每一个单元中的交互过程可表示为方程

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \\
i_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \\
o_t &= \sigma(W_t \cdot [h_{t-1}, x_t] + b_t), \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t, \\
h_t &= o_t * \tanh(C_t),
\end{aligned}
\tag{*}
$$

其中 $\sigma(\cdot)$ 为 sigmoid 函数。前三个方程分别根据前一步的隐藏状态 h_{t-1} 和当前数据输入 x_t，来决定长时记忆 C_t 中遗忘、输入和输出的比例，第四个方程产生待输入记忆的部分，第五个方程执行实际的遗忘和输入过程，最后一个方程执行实际的输出过程。

以上便是经典 LSTM 模型的工作过程及原理。LSTM 还有许多变形，如门控循环单元（Gated Recurrent Unit, GRU）等。这些模型在处理序列数据的任务上都表现出了令人满意的结果。在工程实践中，往往在 t 时刻，利用之前 m 个数据 $x_{t-m}, \cdots, x_{t-1}$ 得到当前的预测值 x_t，并在新的预测值的基础之上，利用 $x_{t-m+1}, \cdots, x_t$ 来预测 $t+1$ 时刻的值 x_{t+1}，以此类推，从而实现长时程的预测过程。

4.2 层次 LSTM 模型

对于风速预测的任务而言，它的另一个难点在于，数据的序列结构并非停留在一个层次上，而是在多个时间尺度上都有体现，即数据的相关性既存在于邻近的样本之间，也存在于间隔了几千个数据的两个样本之间。例如，当前时间的风速不仅与前 15 分钟的状态有关，还和日周期、月周期，甚至季节周期等都有关。因此，我们对该预测任务引入层次结构模型，每个层次处理不同时间分辨率上的序列结构关系，结构如图 5 所示。最底层的模块处理单个样本，越高层的模块处理的时间尺度越大，时间分辨率相应地就越低。

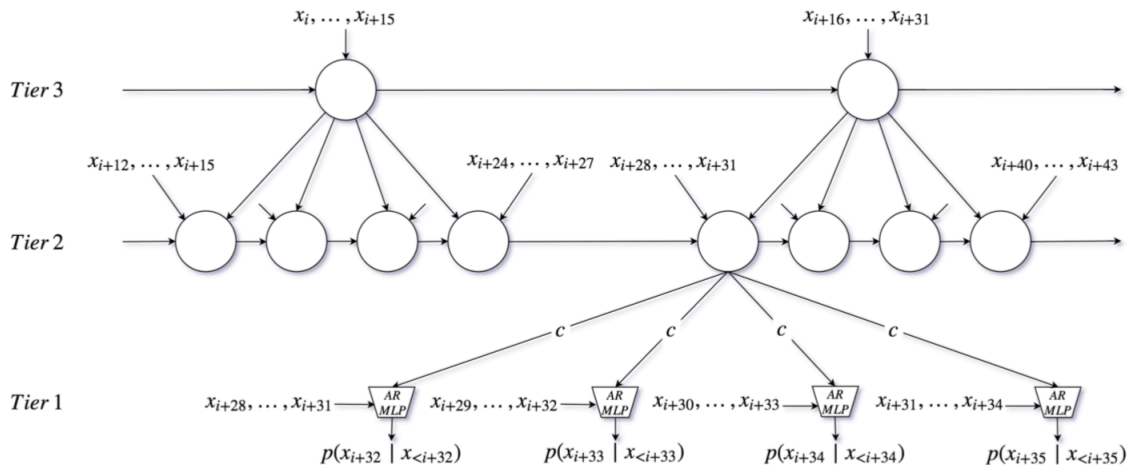


图 5: 层次 LSTM 的结构示意图（上采样率 $r = 4$）

具体而言，假定我们现在有 K 层的 LSTM 模型，那么最顶层（即第 K 层）以最大时间尺度的输入 $f_t^{(K)}$ 作为输入数据，并为下一层的每个模块产生一个条件向量 $c_t^{(K)}$（此步进行上采样，称需要产生的向量个数 r 为上采样率），下一层的模块则同时以稍小时间尺度上的数据 $f_t^{(k)}$ 以及来自上一层的条件向量 $c_t^{(k+1)}$ 共同作为输入，并进行记忆单元的更新，这样逐层向下直至最底层，最底层负责单个样本的处理，其输出作为相应样本的预测值。整体过程可由方程表示为

$$
\begin{aligned}
input_t^{(k)} &= \begin{cases} W_x f_t^{(k)} + c_t^{(k+1)}, & 1 < k < K, \\ f_t^{(K)}, & k = K. \end{cases} \\
h_t^{(k)} &= \mathcal{H}(h_{t-1}^{(k)}, input_t^{(k)}), \\
c_{(t-1)r+j}^{(k)} &= W_j h_t, \quad 1 \le j \le r,
\end{aligned}
$$

其中 $\mathcal{H}$ 为记忆单元更新的过程，这里即指 LSTM 单元。整个网络结构为端到端可训练的，且由于不同的模块处理不同时间层次上的结构关系，所以当模型在为不同尺度上的时序进行抽象时，我们可以有较大的灵活性来分配相应的计算资源。

5 模型效果评估

在实际的建模过程中，由于计算资源的限制，我们最终采用了经典 LSTM 模型来进行预测和评估。输入序列的长度取为 $m = 60$，即利用当前时间点之前 900 分钟内的信息来预测当前的风速。我们期望建立模型，以对

$$x_t = f(x_{t-m}, x_{t-m+1}, \cdots, x_{t-1} \mid \theta)$$

进行回归，其中 θ 为模型的参数。我们对模型添加了 L_2 正则化项，即对方程式 (*) 中的各矩阵的 Frobenius 范数 $\|W_*\|_F^2$ 及偏移向量的 L^2 范数 $\|b_*\|_2^2$，以正则化参数 λ 作为额外惩罚，来控制各系数矩阵使其拥有良好的性态。模型的训练目标为求得模型参数 $\hat{\theta}$ 以最小化经验风险与结构风险，其中经验风险取平均绝对误差（MAE），即求

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{k=1}^{N} |y_k - f(t_k)| + \frac{\lambda}{2} \left(\sum_{*} \|W_*\|_F^2 + \|b_*\|_2^2 \right),$$

对该目标函数使用 AdaGrad 算法进行优化。

在 RNN 系列模型的训练中，梯度消失和梯度爆炸是两种在梯度反向传播时比较容易出现的问题。针对梯度爆炸现象，一方面我们限制了输入序列的长度，以使得梯度能得到有效的传播；另一方面我们采取梯度截断（gradient clipping）策略，即事先设置一个正的阈值 v，一旦当前梯度 $\boldsymbol{g}$ 的范数超过该阈值，就令

$$\boldsymbol{g} \leftarrow v \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|},$$

这样便有 $\|\boldsymbol{g}\| = v$，既保留了梯度原有的方向，又使得梯度不至于爆炸。

模型的具体设置中，LSTM 的最大反向传播截断长度为 $m = 60$，隐藏状态的维数为 $h = 50$，正则化参数 $\lambda = 2.5 \times 10^{-5}$，学习率为 $\eta = 0.1$，mini-batch 大小为 $B = 50$ 个样本。在模型的训练过程中，每经过一个 mini-batch 的训练，便相应地在测试集中选择一个 mini-batch 计算损失函数。测试集上的损失函数也定义为平均绝对误差，即

$$\text{MAE} = \frac{1}{B} \sum_{k=1}^{B} |y_k - f(t_k)|,$$

由此可画出测试误差随迭代次数的变化趋势，如图 6 所示，其中横坐标为训练过的 mini-

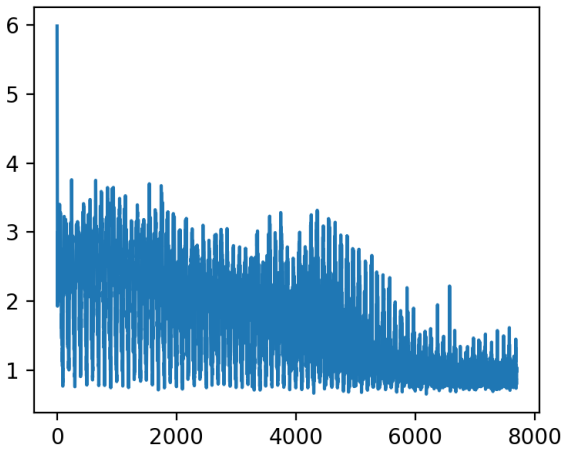


图 6: 测试误差随迭代轮数的变化趋势

batch 总数，纵坐标为测试集上的平均绝对误差。从图中可以看到在 6000 次迭代后，损失函数出现较为明显的下降，且此后的误差均值大致稳定在 1.22 左右，即平均每个数据点上的预测误差为 1.22。对于风速预测这样高难度的任务而言，收敛到这样一个结果是非常令人满意的。

6 结论及展望

对于风速预测，自然风的随机性、间歇性与不可控性大大提高了这一预测任务的难度。我们主要讨论了影响风速的主要因素，以及风速的周期性变化规律。为了把握这一潜在的时序规律，我们采用了在处理序列数据任务上表现非常出色的 LSTM 模型，以当前时间点前的 60 个数据点为基础，进行多尺度的建模，发掘不同层次上的时序抽象关系，且最终的预测结果在插值结果上的拟合效果令人满意，

我们认为，为了更精确地进行风速预测，还需要考虑以下方面的因素，并有针对性地对这些因素进行提高：

(1) 提高数值天气预报的精度。有相关调查表明，风速预测的误差很大一部分来源于天气预报的不准确性，而天气预报的不准确性又恰恰是来自天气的不稳定性，因而天气预报的准确性和风速预测的准确性之间是相辅相成的；

(2) 对气象因子与风速变化的关系进行全面有效的分析。气象信息（包括气压、温度、湿度等）在很大程度上会影响到风速，但风速对这些因素之间的依赖关系较为复杂，可考虑物理建模，或者更为有效的机器学习模拟方法进行系统有效的分析，前提是能够获得较为准确的气象信息；

(3) 考虑对地形地貌的类型与气象信息类型的关系建立数据库，更加全面地考虑影响风速的各个因素，利用相近地形地貌处的气象类型信息辅助当地的风速预测推断，建立更加逼近真实的风电场风速变化模型；

(4) 统一风速预测精度的评价标准，使得各种风速预测方法之间更具有可比性。

在这些工作的基础之上，我们相信，风速预测的精确程度必能有显著性的提高，在理论上对时序关系能有更加深刻的理解，并在实践上创造出最大的经济效益和社会效益。

参考文献

[1] S. Mehri, K. Kumar, Y. Bengio, et al. Sample RNN: An Unconditional End-to-end Neural Audio Generation Model. ICLR 2017 conference paper.

[2] I. Goodfellow and Y. Bengio. Deep Learning. MIT Press, 2016.

[3] C. Olah. Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[4] 王晓称. 风速预测的研究与应用. 华北电力大学.