塔河流域径流量预测技术路线

1. 数据准备与分析

以径流量为基础属性数据, 结合降水量、积雪覆盖度/深度变化量、零度层高度、500百帕温度等可能的影响因素属性来对未来时刻的径流量做预测估计, 尽可能多的提供其他属性字段。

1.1 画图分析: 将径流量在折线图上显示, 直观地查看径流量单个属性走势, 如图1.1.1, 或者将径流量与其他属性分别展示到图上, 可以很直白的看到其他属性与径流量的相关情况, 如图1.1.2。

图1.1.1

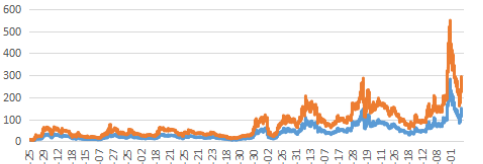


图1.1.2

1.2 主成分分析: 主成分分析就是将大量的彼此可能存在相关关系的变量转换成较少的、彼此不相关的综合指标的一种多元统计方法。这样既可减少变量的应用量，节省算法的运行时间，又使各变量不交叉，便于分析。其主要步骤包括：

* 原始数据标准化处理；
* 计算标准化矩阵的决定系数矩阵；
* 根据决定系数矩阵，得到矩阵的特征值和特征向量；
* 通过计算，得到方差贡献率和累计方差贡献率；
* 确定主成分：即累计方差贡献率大于95%时包含的数据信息总量；
* 计算出各主成分的得分。

1.3 相关性系数分析: 各项属性的相关性程度, 如图1.3.1。



图1.3.1

1. 数据预处理

2.1数据补齐: 保证数据的连续, 采用线性插值, 公式:；

其中，(k, j)开区间表示缺失数据时间范围, k是缺失范围的起点, 是k时刻的径流量, j是缺失范围的止点，是j时刻的径流量, , 表示要插值的时刻缺失的径流量的值。

2.2 如果数据量足够, 可以直接忽略缺失部分。

2.3 离群点检测及数据修复: 数据质量会严重影响预测效果, 因此还要对数据序列进行利群点进行检测和修复.可以采用绝对中位差(Median Absolute Deviation, MAD)来做, 其公式为:



是待检测序列, 是待检测序列的中位数。

2.4 归一化处理: 数据归一化就是把数据经过某种算法处理，使数据的大小限制在一定的范围内，而不改变数据的性质的方法。归一化处理后，使得数据处理起来更加方便、快捷，加快程序的运行速度。这里使用Max - Min 标准化, 公式:, 其中, x 表示原始值, 表示归一化后的结果, 是样本中最小值, 是样本中最大值。

3 数据集制作

为提高预测模型的预测准确度，针对归一化后的径流量数据采用时间滑窗法来构建数据集。利用时间滑窗法对原始数据进行处理可以达到数据维度增加的效果，从而使得预测模型在进行预测时能够提取得到更为丰富的数据特征信息；同时由于径流量数据具有时间关联性（即前后日之间存在一定的延续性），基于时间滑窗法构建的数据集更能反映径流量的变化趋势，从而提高模型的准确度。时间滑窗法的示意图如图 3.1 所示: 滑动窗口以一个步长（Slide）不断向前滑动，窗口的长度固定（Size）。

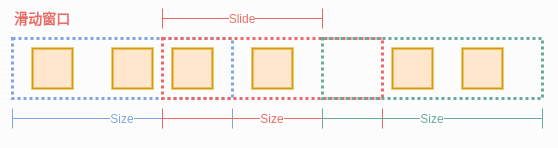


图 3.1

4.模型

4.1. 模型选择: 基于 Bi- Seq2seq 模型的预测框架, 改进的 Bi-Seq2seq 模型，即在Seq2seq模型的编码器中引入双向 LSTM网络（Bi-LSTM）,正反向学习序列数据的内容，掌握全局信息的关联。

4.2. Bi-LSTM原理: Bi-LSTM是由LSTM演变而来。

4.3. LSTM其单元结构如图4.3.1, LSTM 拥有三个门, 即遗忘门、输入门、输出门。

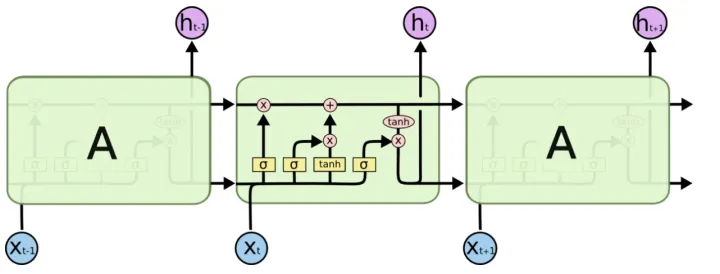


图4.3.1

遗忘门:如图4.3.2, 该门决定丢弃的信息, 计算和，通过激活函数后, 会输出一个在0到1之间的值, 再将这个值作用到每个细胞的状态去。1 表示“完全保留”，0 表示“完全舍弃”, 计算公式为:



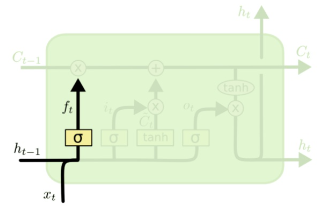


图4.3.2

输入门: 如图4.3.3, 该门确定更新的信息。 这里包含两个部分, 第一，称 “输入门” 决定要输入的东西。然后，一个 创建一个新的候选值向量，会被加入到细胞状态中。计算公式为:



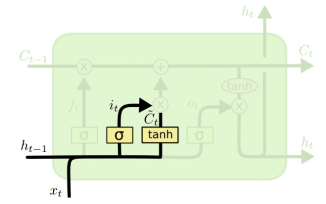


图4.3.3

接下来则是更新细胞状态, 如图4.3.4, 将更新为, 计算公式为:



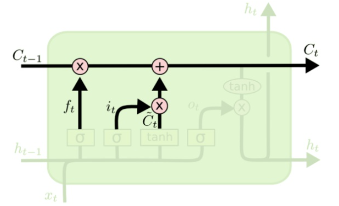


图4.3.4

输出门: 如图4.3.5, 该门控制输出信息。计算公式为:



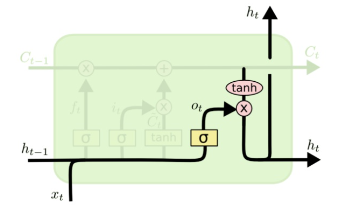


图4.3.5

4.4. Bi-LSTM则是由前向的LSTM与后向的LSTM结合成BiLSTM。Bi-LSTM 神经网络结构模型分为 2 个独立的LSTM，输入序列分别以正序和逆序输入至 2 个LSTM 神经网络进行特征提取，将 2个LSTM输出结果进行拼接后形成最终的结果。Bi-LSTM 的模型设计理念是使 t 时刻所获得特征数据同时拥有过去和将来之间的信息，实验证明，这种神经网络结构模型对文本特征提取效率和性能要优于单个 LSTM 结构模型。值得一提的是，Bi-LSTM 中的 2 个 LSTM 神经网络参数是相互独立的. 如图4.4.1, 其表达式可以写为:



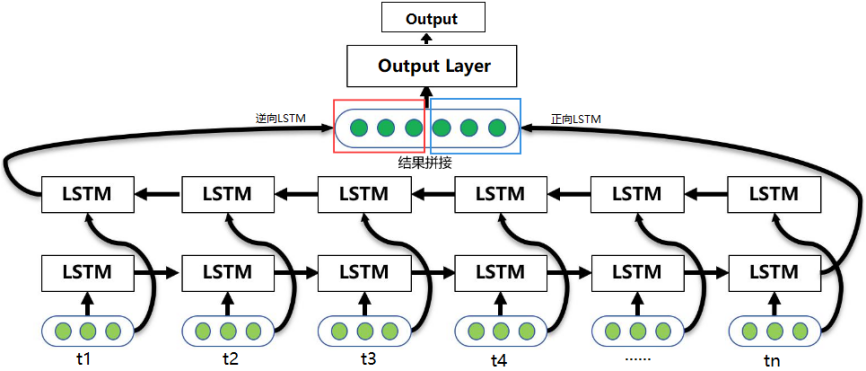


图4.4.1

4.5. Bi- Seq2seq原理: 包含了Encoder-Decoder两个模块, 如图4.5.1所示, 在 Bi-Seq2seq 模型的编码器(Encoder)采用双向长短期记忆网络(Bi-directional Long Short-Term Memory, Bi-LSTM)提取特征， 高质量地学习正反向序列数据间的关联信息, 再将正反向数据拼接， 减少数据信息遗忘. 解码器(Decoder)采用 LSTM 网络输出未来任意步数的值, 实现任意长度的输入序列映射到任意长度的输出序列, 同时保留其序列之间的依赖性.

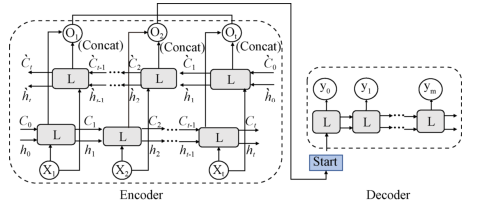


图4.5.1

5.输出

5.1 评估指标: 假定m是预测目标总个数, 是实际值，是预测值:

MAE(Mean Absolute Error) 平均绝对误差, 计算实际值与预测值的距离绝对值做评估, 其公式为:

；

MSE(Mean Square Error)均方误差, 计算实际值与预测值的距离平方做评估, 其公式为:

；

RMSE(Root Mean Square Error)均方根误差, 即MSE开放的结果, 其公式为:

；

MAPE(Mean Absolute Percentage Error)平均绝对百分比误差, 不仅考虑预测值与真实值的偏差, 还考虑了偏差与真实值之间的比例，此指标对相对误差敏感，不会因目标变量的全局缩放而改变, 其公式为:

；

5.2 输出结果画图展示, 可以很直观的看到预测的效果, 如图5.2.1。

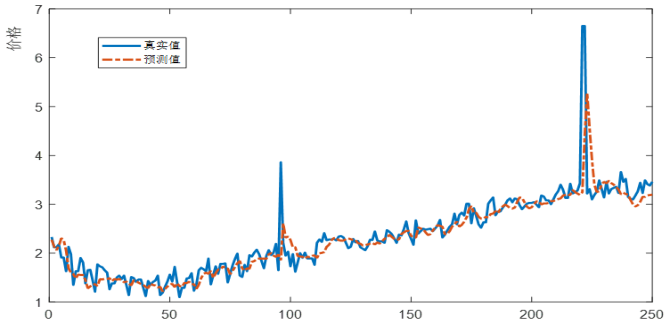


图5.2.1