Consider from the problem itself, two of our teammates who is from the department of mathematics selected the factors which affects the river water level most from all the characteristics through correlation analysis from the perspective of hydrogeology. These factors contains hydrology factors, geology factors and environment factors which includes precipitation, river bed properties etc.

Because our purpose is to predict water level, and in data science, only historical data can predict the future. So we model these data with periodic changes as time series prediction problem.

Next is the feature engineering. We transform the existing features into more effective features. First of all, we differentiate the data of the discharge and water level of different stations. Because all the external factors affect the change of water level, not the water level itself.

Then, we select the most relevant historical data of the stations through correlation analysis, and use them to represent the hydrological information.

Because different kinds of precipitation has different time delay, we classify it as liquid precipitation, new snow and old snow for processing.

In addition, compared with the traditional method of directly processing meteorological data, we creatively use Kriging and cokrigging interpolation to transform the data of scattered meteorological stations into image data for further prediction which generate three images contains precipitation, new snow and old snow.

and we use Kriging ans cokrigging interpolation to transform the discrete data into continuous precipitation distribution image.

Similarly, combining snow data and terrain data, we use Cokriging interpolation to generate new and old snow distribution map

After getting the above preprocessed feature data, we use a model which is a combination of CNN and LSTM to predict the water level. The structure of CNN is responsible for one-dimensional convolution of the historical information of each station and integrate our preprocessed data. And the structure of LSTM is responsible for the prediction of time series using the fusion information generate by CNN.

从问题本身考虑,我们两位数学系的同学在水文地质学角度, 通过相关分析从所有特征中选取了影响河流水位的因素which contains 水文(插图)factors, 地质(c)factors, 环境(c)factors, which includes

降水,河床性质,以及支干流关系. 而具体到莱茵河的问题, 通过…资料,我们知道莱茵河的地质情况在50年内的变化甚微,所以我们认为可以忽略.(来源)

由于我们的目的是预测水位,而在数据科学中,能对未来进行预测的数据只有历史数据.所以我们把这些有周期性变化的数据建模为时间序列预测问题.

接下来是特征工程,我们将已有特征设计为更有效的特征.首先, 我们对不同测站的流量和水位进行向前差分(图),因为所有的外界影响因素影响的是水位的变化值而不是水位值本身. 然后,我们通过相关分析,选取出最相关的测站的历史数据,(测站分布图) 并用他们表示水文信息. 其次, Because different kinds of 降水 has different 时间延后性 ,所以我们将其归为液态降水,新积雪,旧积雪三类进行处理(图).

此外,相较传统的直接处理气象数据的方法, 我们创新性地分散的气象站的数据转换为了图像数据,我们使用了使用krigging 插值将离散数据转化为连续的降水量分布图像. 同理,结合积雪数据和地形数据我们使用cokrigging 插值生成新旧积雪分布图.(图)

得到以上预处理后的特征数据,我们利用CNN+LSTM(图)模型进行预测, 其中CNN的结构负责将每个测站的历史信息进行一维卷积,把我们预处理好的数据(可视化)融合在一起 , LSTM的结构负责利用融合的信息进行时间序列预测以得到我们最后的结果.