天津大学

统计数据分析专题报告

学　　院　国际工程师学院

年　　级　2018

组　　长　后永胜

小组成员　刘虹　李亚蓉

段义海　杜洪才

2019 年 11 月 4 日 星期一

基于监督学习的多变量水力发电站短期来水预测

后永胜*, 刘虹, 李亚蓉, 段义海, 杜洪才

Tianjin International Engineering Institute, Tianjin University, Nankai, Tianjin, China, 300072

Abstract

水力发电具有清洁、无污染、廉价、可再生的优点，因此水电生产成为我国当前重要的发电方式。水电站的入库流量受到降雨量、上游水库出水量的影响，是一个很难被确定的随机变化量。然而对入库流量变化情况的掌握又恰恰是指定发电计划及其它用水调度的基础。传统的统计学习方法 ARIMA，只能考虑单变量的影响，通过此方法建模得到的模型鲁棒性较差，不能充分考虑到雨量和上游出水量的影响。本文通过充分挖掘影响水库进水量的多变量的有效信息，构造有效的特征空间，利用贝叶斯岭回归方法、基于树的提升方法和 Elman 神经网络进行建模预测，最终我们的模型取得了很好的单时间步水库进水量预测的效果。

1. 引言

能源问题是 21 世纪全世界最关注的问题之一，“发展清洁能源和可再生能源”都被各国政府置于高度的发展战略地位。水能作为电能生产的一次能源，具有清洁、无污染、廉价、可再生等特点，在社会经济发展中有着极其重要的作用。因此，水电生产成为我国政府大力支持和发展的产业之一。

水电站的入库流量受到降雨量、蒸发量、上游电站运行情况等因素的制约，是一个不能被确定的随机变化量。但它对水电站电能生产有着重要的影响作用，对入库流量变化情况的掌握又恰恰是制定发电计划、进行电网调度的基础。

本文的研究目标是根据上下游水库小时级历史流量数据，结合前期水文、气象等资料，实现对下游水库未来进水量的科学预测。

传统的统计学习方法 ARIMA，模型只考虑来水流量一个时间序列，而忽略了上游流量序列的影响。实际上，在大部分情况下，上游出库流量是影响下游入库流量的一个关键因素，通过此方法建模得到的模型鲁棒性较差，不能充分考虑到雨量和

*Corresponding author: houys@tju.edu.cn

2019 年 11 月 6 日

上游出水量的影响。本文通过充分挖掘影响水库进水量的多变量的有效信息，构造有效的特征空间，利用贝叶斯岭回归方法、基于树的提升方法和 Elman 神经网络进行建模预测，最终我们的模型取得了很好的单时间步水库进水量预测的效果。

2. 解决方法

2.1. 特征工程

本文我们选择的方法都是有监督的学习模型，为了最大限度地从原始数据中提取特征以供算法和模型使用，我们进行了大量的工作来构建特征空间。主要包括数据清洗，数据特征抽取以及特征降维等。接下来我们将详细讲解进行特征工程的主要思路。

2.1.1. 数据概览

首先，画出水库水量信息和雨量信息的三年的数据趋势曲线，其中这五列数据分别代表的是上游水库出水量，下游水库进水量，以及三个降雨信息观测点的雨量信息。数据的时间粒度是小时。如图1所示。可以看出，降雨量具有明显的周期特征，六月份，七月份和八月份的降雨量最多；而水量趋势无明显的周期特征。

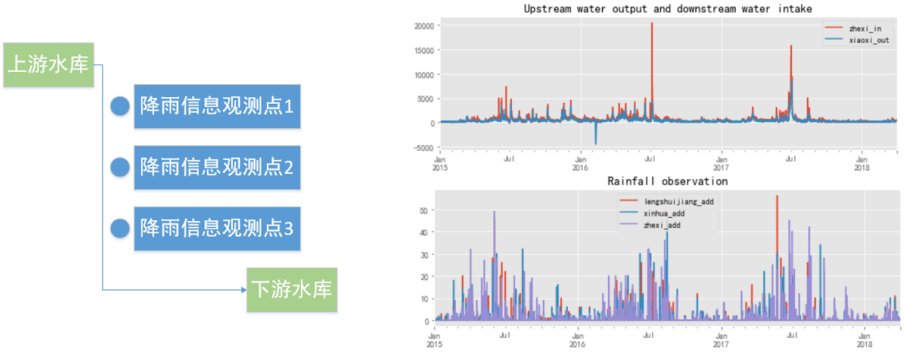

图 1: 数据整体趋势。数据具体含义 (左图)，水量整体趋势 (右图上)，雨量整体趋势 (右图下)。

2.1.2. 数据分布

第二步，画出数据分布图，如图2所示。上游出水量和下游进水量数值 99% 以上在 0-2000 内，雨量信息大致在 0-10 以内。

2.1.3. 数据清洗

主要包含空值处理和异常点的处理。其中空值处理是用纵向用缺失值前面的值替换缺失值。在分布图中可以看到上游水库出水量有一些负值，这是由于水库设置了一个水位线，低于此水位线为负，高于为正。在此处理将这些负值重新赋值为 0 。

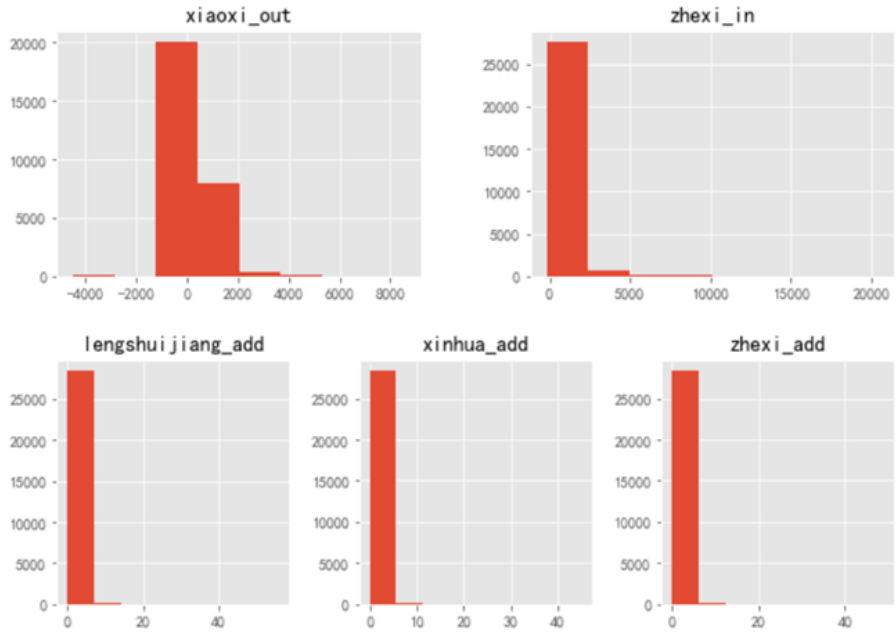

图 2: 数据的历史分布图。水量的分布 (上图)，雨量的分布 (下图)。

2.1.4. 特征构建

提取重要的数据特征，主要包括：

(1) 最近 24 小时的水量平均值特征

(2) 最近 24 小时的降雨量总和特征，以及降雨强度类别特征

(3) 下游水库最近 period 个进水量值

(4) 上游水库 period 个时刻的出水量，经过 offset 个小时的延迟，影响到当前时刻

其中，period 和 offset 是超参数，分别代表着上下游历史水量对下游水库进水量的影响区域，以及上游水库出水量经过一定的时间到达下游水库的时延。图3可以清楚展示 period 和 offset 的含义。

2.1.5. 超参数优化

我们刚刚提出了两个超参数 offset 和 period，这两个参数会影响到我们构建的特征空间的维度，下面我们将利用贝叶斯优化来寻找最佳的 offset 和 period。设定搜索空间为 $offset \in [36, 60]$, $period \in [5, 25]$。家下来设定待优化的黑盒函数，

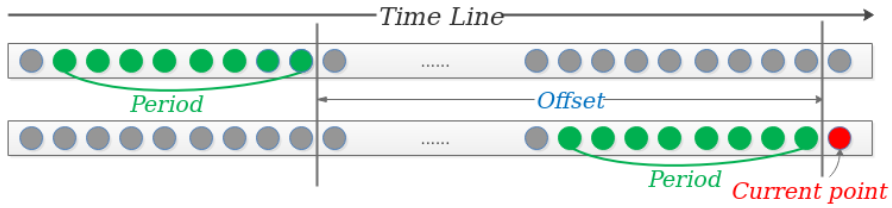

图 3: 水量历史信息特征。offset 表示上游到达下游水库的时间延迟，period 表示影响的区域点数。

首先拿出部分数据作为训练数据，将数据按照不同的 offset 和 period 值构建特征空间，然后将数据划分为训练集和验证集，然后利用贝叶斯岭回归算法进行建模预测，最后的目标是最小化真实值与预测值的均方误差。贝叶斯优化的结果如图4所示。我们进行了 150 轮搜索，每次随机在五个点附近依概率采样，最后得到的结果 $offset = 59, period = 24$, 最小的均方误差为 22934.69。

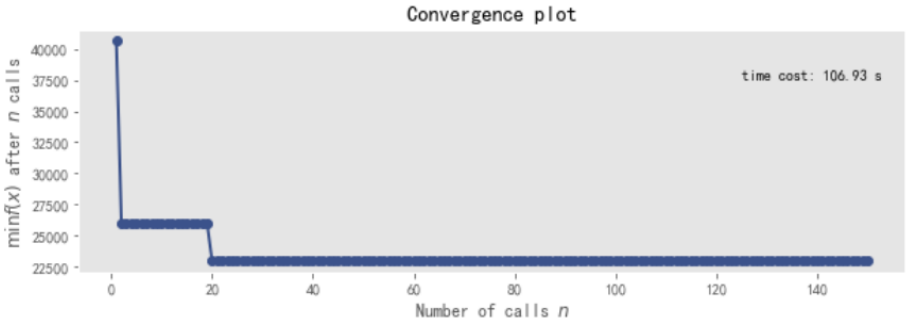

图 4: 贝叶斯优化的结果。搜索 150 轮，最小的均方误差为 22934.69

2.1.6. 特征空间有效性分析

目前为止，我们已经构建好了特征空间，但是这些特征空间是否有效，尚不可知。所以，在此我们进行简单的相关性分析。计算构造的特征与目标的相关系数，以此大致推断构造的特征的有效性。我们由图5可以看出，黄色和绿色具有较大的相关系数，黑色的具有较小的相关系数。我们去掉这些具有弱相关性的特征。我们可以继续看出相关度最大的特征 top10 中，很大一部分是上游出水量历史信息，这说明在现实中上游出水量对下游水库的进水量在很大程度上有着决定性的影响。

2.1.7. 特征空间降维

现在我们得到的特征空间的信息冗余度比较高，而且特征维度越高，模型也会相对应的变得复杂。我们当前的数据量不大，同样的数据量的条件下，模型相对简单的会

图 5: 特征与目标的相关系数。相关度热度图 (左图)，相关性最大的特征 top10(右图)。

训练的更好。所以我们要对特征空间降维。在降维之前，我们首先利用 BayesianRidge 方法对当前得到的特征空间进行建模测试，然后利用 PCA 将特征空间降到 20 维，最后利用 VAE 方法将特征空间降到 20 维，得到的结果如图6所示。

至此，我们为算法模型提取到了有效且低维度高表征的特征空间，下面我们将利用它们建立监督学习的模型进行预测分析。不过在此之前我们先利用了传统的统计学习方法 ARIMA 方法进行建模。

2.2. 差分整合移动平均自回归

2.2.1. 模型简介

ARIMA 模型（Autoregressive Integrated Moving Average model），差分移动平均自回归模型，时间序列预测分析方法之一。自回归模型描述当前值与历史值之间的关系，用变量自身的历史时间数据对自身进行预测。自回归模型必须满足平稳性的要求。移动平均模型关注的是自回归模型中的误差项的累加，移动平均法能有效地消除预测中的随机波动。模型公式:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-1} + \epsilon + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

模型参数：p, d, q, 其中参数 p 代表预测模型中采用的时序数据本身的滞后数；参数 d 代表时序数据需要进行几阶差分化，才是稳定的；参数 q 代表预测模型中采用的预测误差的滞后数。

2.2.2. ARIMA 建模

(1). 模型识别和定阶

首先观察数据是否是平稳序列，对于非平稳时间序列进行 d 阶差分运算，化为平稳时间序列; 分析自相关图和偏自相关图，得到阶层 p 和阶数 q，自相关函数 ACF 描述时间序列观测值与其过去的观测值之间的线性相关性，公式：

$$ACF(k) = \rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)}$$

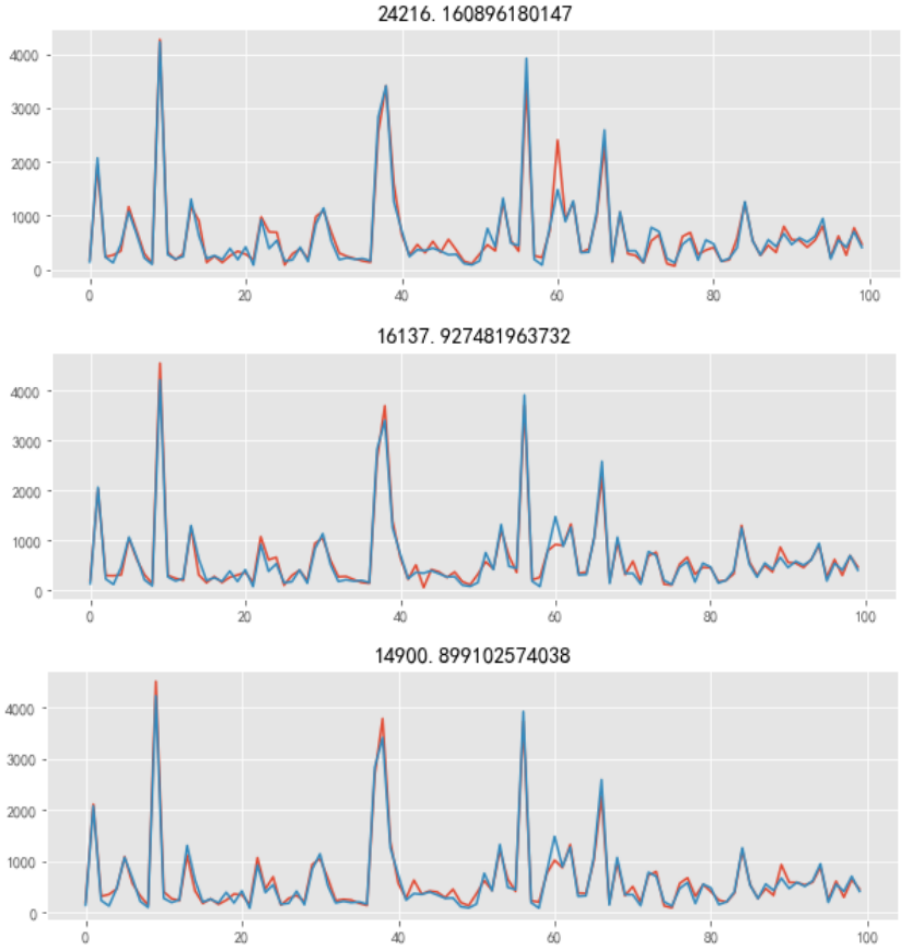

图 6: 特征空间降维的测试结果。原始特征空间预测结果 (上),PCA 降维后预测结果 (中),VAE 降维后预测结果 (下)。

偏自相关函数 PACF 描述在给定中间观测值的条件下，时间序列观测值与过去的观测值之间的线性相关性。拖尾指序列以指数率单调递减或震荡衰减，而截尾指序列从某个时点变得非常小 p，q 的确定基于如图7所示:

图 7: 参数 p 和 q 的确定规则。

(2). 参数估计方法

通过拖尾和截尾对模型进行定阶的方法，往往具有很强的主观性。在相同的预测误差情况下，根据奥斯卡姆剃刀准则，模型越小是越好的。那么，平衡预测误差和参数个数，我们可以根据信息准则函数法，来确定模型的阶数。预测误差通常用平方误差即残差平方和来表示。常用的信息准则函数法有 AIC 准则和 BIC 准则。AIC 准则：全称是最小化信息量准则 (Akaike Information Criterion), 计算公式为 $AIC = 2 * (num\_params) - 2\ln(MLE)$，MLE 为极大似然函数。BIC 准则：AIC 准则存在一定的不足之处。当样本容量很大时，在 AIC 准则中拟合误差提供的信息就要受到样本容量的放大，而参数个数的惩罚因子却和样本容量没关系（一直是 2），因此当样本容量很大时，使用 AIC 准则选择的模型不收敛与真实模型，它通常比真实模型所含的未知参数个数要多。BIC（Bayesian InformationCriterion）贝叶斯信息准则弥补了 AIC 的不足，计算公式为 $\ln(n) * num_params - 2\ln(MLE)$，n 是样本容量。

(2). 模型校验

检验参数估计的显著性 (t 检验)，检验残差序列的随机性，即残差之间是独立的。

(3). 模型预测

预测主要有两个函数，一个是 predict 函数，一个是 forecast 函数，predict 中进行预测的时间段必须在我们训练 ARIMA 模型的数据中，forecast 则是对训练数据集末尾下一个时间段的值进行预估。

2.2.3. 结果分析与评价

我们选取下游水库进水量的一段数据机型建模预测，得到的结果如图**??**所示。该模型的结果是预测未来 10 小时水库进水量的值，红色区域是预测的最大值和最小

值，蓝色曲线是预测结果，红色曲线是真实值。通过对比发现，预测时间越长，结果偏差越大，所以该模型不适合做长时间的序列预测。且该模型只考虑到了下游水库的历史进水量，没能考虑到降雨量和上游水库的出水量信息，模型鲁棒性较差。下面我们将利用监督学习的方法构建模型，加入这些变量的影响，这样可以使得到的模型具备更好的鲁棒性。

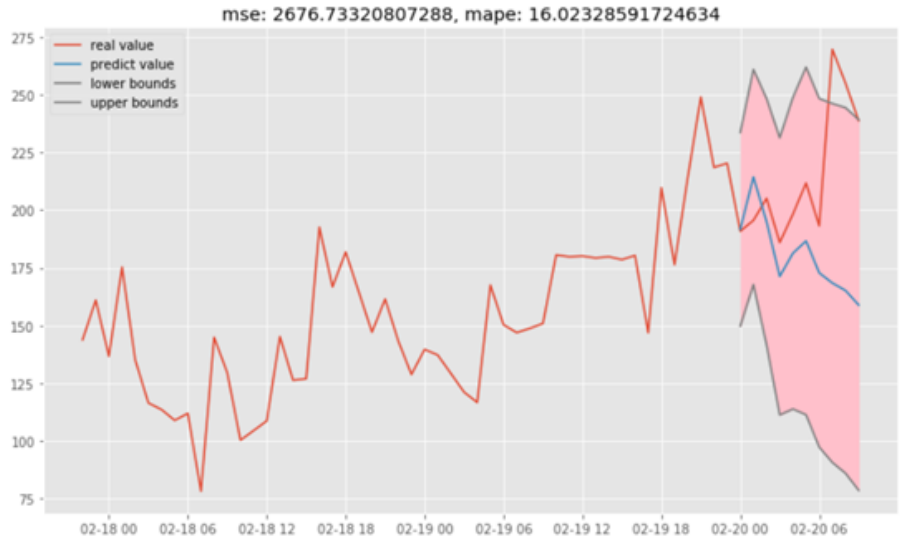

图 8: ARIMA 模型简单预测结果。

2.3. 贝叶斯岭回归

在最大似然估计中很难决定模型的复杂程度，岭回归加入的惩罚参数可以解决这个问题。贝叶斯线性回归将线性模型的参数视为随机变量，并通过模型参数的先验计算其后验。贝叶斯线性回归可以使用数值方法求解，在一定条件下，也可得到解析型式的后验或其有关统计量。再贝叶斯回归方法中加入岭回归的惩罚项，得到的贝叶斯岭回归算法，对于数值回归分析问题的解决具有很大作用。

2.3.1. Gamma 分布

统计学的一种连续概率函数，是概率统计中一种非常重要的分布。“指数分布”和“$\gamma 2$ 分布”都是伽马分布的特例。Gamma 分布中的参数 称为形状参数（shape parameter）， 称为逆尺度参数假设随机变量 γ 为等到第 α 件事发生所需之等候时间, 密度函数为：

$$f(x,\beta,\alpha)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x},x>0$$

。

2.3.2. 贝叶斯岭回归

贝叶斯岭回归是估计回归问题的概率模型，系数 w 的先验由球面高斯给出：$p(w \mid \lambda) = N(w \mid 0, \lambda^{-1} I_p)$。$\alpha$ 和 λ 的先验为 gamma 分布，这是高斯精度的共轭先验。所得模型称为贝叶斯岭回归。w, α, λ 在模型的拟合过程中被共同估计，在似然函数的基础上引入参数的先验分布，借助似然函数和先验分布去最大化参数的后验概率，即

$$p(w \mid x, t, \alpha, \beta) \propto p(t \mid x, w, \beta) p(w \mid \alpha)$$

2.3.3. 评估指标

MAPE: 平均绝对百分比误差

$$MAPE = \sum_{i=1}^{n} \mid \frac{\widehat{y_i} - y_i}{y_i} \mid * \frac{100}{n}$$

不仅仅考虑预测值与真实值的误差，还考虑了误差与真实值之间的比例，在统计领域是一个预测准确性的衡量指标。MSE: 均方误差

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\widehat{y_i} - y_i)^2$$

MSE 可以放大预测偏差较大的值，可以比较不同预测模型的稳定性，应用场景相对多一点。

2.3.4. 结果分析与评价

我们利用 2015 年的数据训练模型，用 2016 年的数据测试模型，得到的结果如图9所示。由图中可以看出，预测的结果大致拟合原始数据，由此说明我们的模型取得了很好的预测效果。

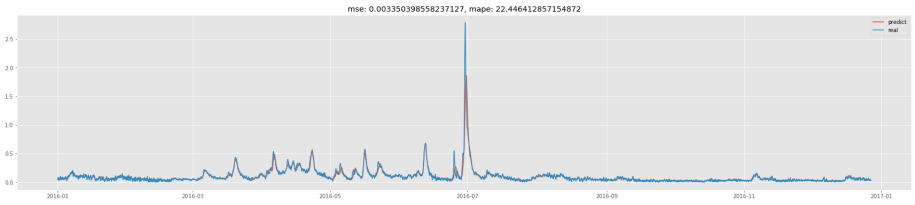

图 9: 贝叶斯岭回归方法。2015 年的数据预测的 2016 年的数据。

图10展示了贝叶斯岭回归模型的学习曲线, 由图中可以看出模型在训练集和测试集的评估分数最后趋与一致，说明我们的模型处于良好的学习状态，并未出现过拟合或者欠拟合的情况。

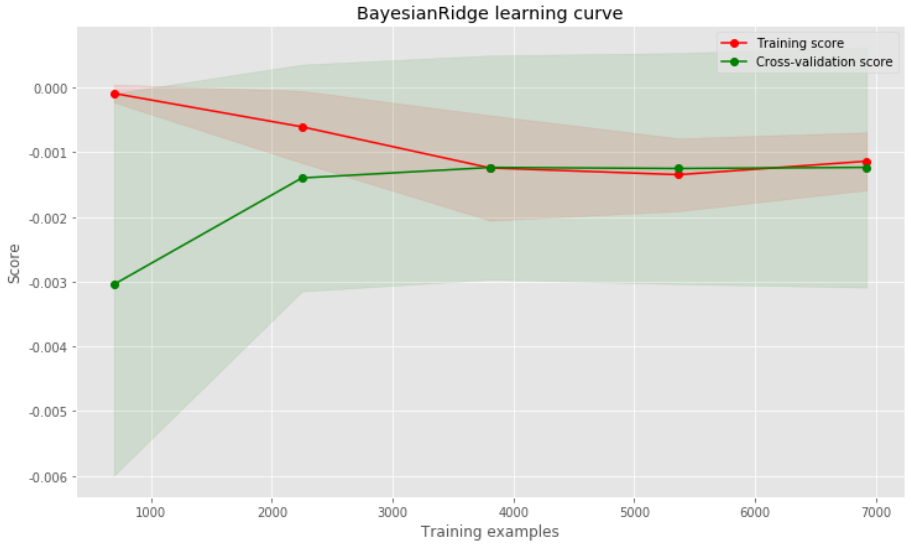

图 10: 贝叶斯岭回归方法。学习曲线。

本节中我们利用了贝叶斯岭回归方法建模预测，该方法是线性模型，下面我们将采用两种非线性的方法建模分析：基于树的随机梯度提升算法和多层感知机 Elman 神经网络。

2.4. 基于树的提升方法

目前有三大主流的基于树的提升方法，它们分别是 xgboost,catboost 和 lightgbm，这些方法在诸如 kaggle 的数据竞赛中广泛使用，它们往往能取得较好的预测效果。所以我们也利用这三种方法进行了建模分析。

2.4.1. 决策树简介

决策树的决策过程就是从根节点开始一步步走到叶子节点。决策树是一种简单高效并且具有强解释性的模型，广泛应用于数据分析领域。其本质是一颗由多个判断节点组成的树，在使用模型进行预测时，根据输入参数依次在各个判断节点进行判断游走，最后到叶子节点即为预测结果。树形模型与线性模型的最大不同指出在于：树形模型是一个一个特征进行处理，之前线性模型是所有特征给予权重相加得到一个新的值，另外树形模型因为均有很好的可解释性，因此在目前生产生活中应用很广泛。

2.4.2. GBDT 算法原理

GBDT 的原理是，首先使用训练集和样本真值（即标准答案）训练一棵树，然后使用这棵树预测训练集，得到每个样本的预测值，由于预测值与真值存在偏差，所

以二者相减可以得到“残差”。接下来训练第二棵树，此时不再使用真值，而是使用残差作为标准答案。两棵树训练完成后，可以再次得到每个样本的残差，然后进一步训练第三棵树，以此类推。树的总棵数可以人为指定，也可以监控某些指标（例如验证集上的误差）来停止训练。在预测新样本时，每棵树都会有一个输出值，将这些输出值相加，即得到样本最终的预测值。xgboost[1],catboost[2] 和 lightgbm[3] 是基于树的随机梯度提升算法, 它们在训练样本量有限、所需训练时间较短、缺乏调参知识等场景依然有其不可或缺的优势。

2.4.3. 结果分析与评价

利用这三个提升算法，我们利用 2015 年的数据训练模型，预测 2016 年的数据，得到的结果如图11所示。

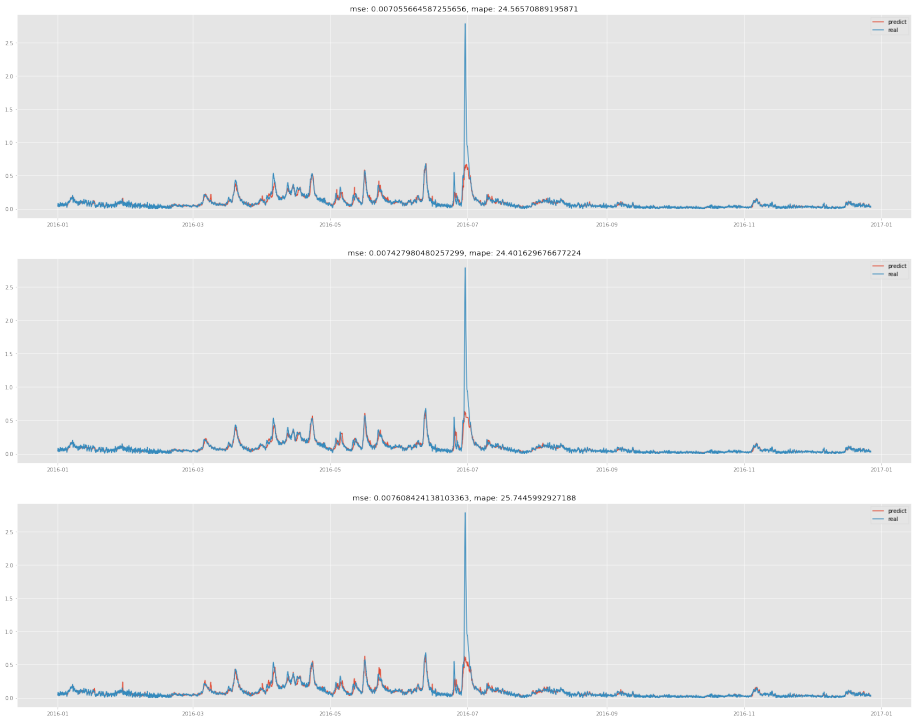

图 11: 三种 Boosting 算法的结果。xgboost(上)，catboost(中),lightgbm(下)。

由图11可见，这三个提升算法的结果相差不大，但是与之前的贝叶斯岭回归方法的结果相比 ($mse = 0.00335, mape = 22.45$), 显得稍差一筹。主要原因可能是这三个 boosting 算法模型的参数过多，模型之中的超参数调优会花费很大的算力和时间，此时模型并未处于最优参数状态；另外，训练数据量可能不大，通过模型的学习曲

线12可以看出，此时模型处于过拟合的状态，因为验证集的 score 跟训练集的 score 有些差距，并没有重合到一起。

图 12: Xgboost 算法模型的学习曲线。

下面我们将用另外一种非线性模型：Elman 神经网络进行建模分析。

2.5. Elman 神经网络

2.5.1. 模型简介

ELMAN 神经网络是在 1990 年首先针对语音处理问题而提出来的，是一种典型的动态递归网络。ELMAN 网络可以看作是一个具有局部记忆单元和局部反馈连接的递归神经网络。它在 BP 网络基本结构的基础上，在隐含层增加一个反馈连接层，作为一步延时算子，达到记忆的目的。模型结构如图13所示。它的结构包括输入层、隐含层、输出层，其连接权值可以进行学习修正；每一个隐含层节点都有一个与之对应的反馈连接结点；反馈连接结点用来记忆前一时刻的输出值，其连接权值是固定的。反馈连接层的作用是记忆将上一个时刻的隐层状态连同当前时刻的网络输入一起作为隐层的输入, 相当于状态递归。隐含层的传递函数为某种非线性函数, 一般为 Sigmoid 函数, 输出层为线性函数, 反馈连接层也为线性函数。

Elman 神经网络的数学表示如下：

$$y(k) = g(w_3 x(k)) \tag{1}$$

$$x(k) = f(w_1 x_c(k) + w_2(u(k-1))) \tag{2}$$

$$x_c(k) = x(k-1) \tag{3}$$

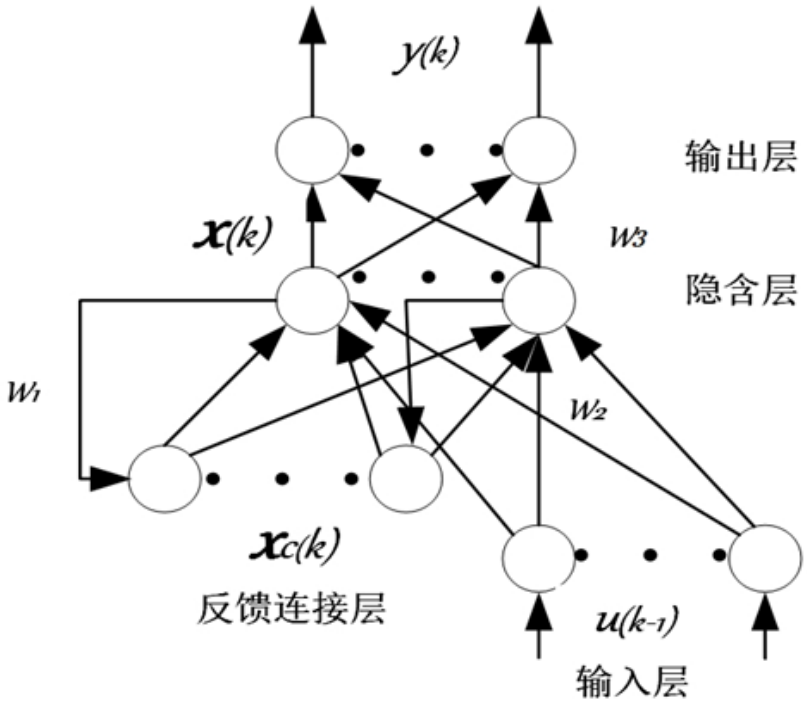

图 13: Elman 神经网络模型结构。

其中 y 为 m 维输出节点向量；x 为 n 维中间层节点单元向量；u 为 r 维输入向量；x_c 为 n 维反馈状态向量；w_3 为中间层到输出层连接权值；w_2 为输入层到中间层连接权值；w_1 为反馈连接层到中间层连接权值；g 为输出神经元的传递函数；f 为中间层神经元的传递函数。

2.5.2. 结果分析与评价

同样我们用 2015 年的数据训练模型，2016 年的数据测试模型。最终利用 Elman 神经网络得到的结果如图14所示。

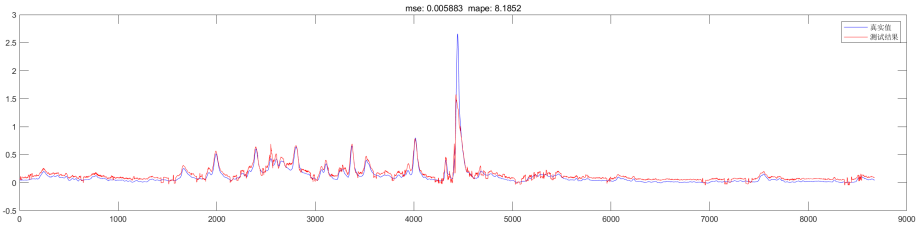

图 14: Elman 神经网络模型的测试结果。

由图14可以看出，结果的均方误差的值在前两种方法之间，但是取得了最小的平均绝对误差值。

3. 结果分析

实验中我们的主要评价指标是均方误差 (mean squared error)，对比三者的结果15可以看出，贝叶斯岭回归方法得到了最小的 mse 值，模型取得了好的效果。其次是 Elman 神经网络模型，最后是 Boosting 方法。其中 Elman 神经网络模型取得了远小于其它二者的 mape 值。

图 15: 三种监督学习算法的结果。贝叶斯岭回归 (上)，Boosting 方法 (中)，Elman 神经网络 (下)。

贝叶斯岭回归方法加入了岭回归方法的正则项，可以有效控制模型的复杂度，使得模型能很好应对不同大小的训练集；对于大的数据集，增加模型的复杂度以增加模型性能，而对于小的数据集，减少模型的复杂度以防止过拟合的风险，这就决定了贝叶斯岭回归方法的优越性能，简单高效。而基于树的随机梯度提升算法基于树模型，可解释行比较强，但是对于训练数据集的大小有一定要求；另外，它们调参工作比较费时，对于算力的要求也比较高。最后，Elman 神经网络模型，虽然只是三层的网络结构，但是由于在网络加入了负反馈机制，使得模型能够“记住”前面时刻的影响，从而达到了较好的预测效果。

4. 结论

ARIMA 方法的本质是单变量的回归预测，不能考虑到其它相关变量的影响，因此鲁棒性较差。而另外三种方法是基于监督学习的方法，可以通过相关变量的特征组合，构造有效的特征空间，进而模型通过学习，得到一个较好的预测结果。其中，贝叶斯岭回归方法取得了最好的单步预测效果。

References

[1] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, CoRR abs/1603.02754 (2016). URL: http://arxiv.org/abs/1603.02754. arXiv:1603.02754.

[2] A. V. Dorogush, V. Ershov, A. Gulin, Catboost: gradient boosting with categorical features support, CoRR abs/1810.11363 (2018). URL: http://arxiv.org/abs/1810.11363. arXiv:1810.11363.

[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 3146–3154. URL: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

附录 A. 代码

所有的代码均已上传到 Github 仓库：https://github.com/mmmmmmiracle/sda_project。代码格式为 jupyter notebook 的.ipynb 文件，python 脚本.py 文件。