Acea Smart Water Analytics

ECE 9309B/9039B Machine Learning: From Theory to Applications Progress Report Information

Xinze Li
Department of ECE
Western University
London, Canada
xli2966@uwo.ca

Qiaomei Han
Department of ECE
Western University
London, Canada
qhan42@uwo.ca

Peizhi Yu
Department of ECE
Western University
London, Canada
pyu83@uwo.ca

***Abstract*—This document is a Progress report for Acea Smart Water Analytics. The article is included the introduction, related work, description of dataset, Methodology, and results analysis sections. The main models for analysis include Multi-layer Perceptron (MLP) neural network, Long and Short-Term Memory (LSTM), Support Vector Regression (SVR) and Gated Recurrent Unit.**

***Index Terms*—Water Analytics, MLP, LSTM, SVR, GRU**

I. Introduction

This project aims to help Acea Group preserve precious waterbodies. As it is easy to imagine, a water supply company struggles with the need to forecast the water level in a waterbody (water spring, lake, river, or aquifer) to handle daily consumption. It is critical to understand total availability in order to preserve water. During fall and winter waterbodies are refilled, but during spring and summer they start to drain. To help preserve the health of these waterbodies it is important to predict the most efficient water availability, in terms of the depth to groundwater, flow rate, hydrometry, etc.

Features can vary based on different types of waterbodies. For instance, the hydrometry of the Arno River as well as the lake level and flow rate of the Bilancino Lake can be predicted using several machine learning methods.

II. Related Work

For preserving water resources, this project aims to analyze the the availability of the waterbodies, which needs to forecast the water level in a waterbody (water spring, lake, river, or aquifer) to handle daily consumption. As the availability is related to some time-dependent indexes, we will focus on the time-series analysis algorithms.

The LSTM network is proposed by Hochreiter and Schmidhuber [1], where the storage blocks replace the traditionally hidden layer neurons. This block prevents the disappearance and explosion of gradients during long-term training. Therefore, LSTM enables to process long-term sequences. Many LSTM-based deep learning models have been successfully applied to predict time-series data. However, achieving reliable and accurate prediction in complex, nonlinear, and time-varying conditions remains a challenge.

The Support Vector Machine has been tried for this project. However, the model results are unexpected. In classification problems, SVMs are well-known. However, the use of SVMs in regression is less well documented. SVR models are the name for these types of models. Thus, the SVR model has been cancelled once and changed to GRU. Finally, after group discussion, the SVR will be use again in this project.

III. Dataset

The dataset is about the Acea Smart Water Analytics provided from Kaggle, https://www.kaggle.com/c/acea-water-prediction/data [3]. There are nine different sub-datasets, completely independent and not linked to each other. Each sub-dataset can represent a different kind of waterbody. As each waterbody is different from the other, the related features as well are different from each other. To simplify, we will consider the information of two sub-datasets and make predictions based on their different features, which is the Arno River and the Bilancino lake.

The availability of Arno River is evaluated by checking the hydrometric level of the river at the section of Nave di Rosano. In this sub-dataset, there are 16 features, which are the rainfalls in 14 locations, the temperature and the hydrometry. Bilancino lake is an artificial lake. During the winter months, the lake is filled up and then, during the summer months, the water of the lake is poured into the another waterbodes. In this sub-dataset, there are 8 features, which are the rainfalls in 6 locations, the lake level and the flow rate.

Therefore, the hydrometry of the Arno River as well as the lake level and flow rate of the Bilancino Lake will be predicted using Multilayer Perceptron neural network, Long short-term Memory, and SVR in this project. The detailed descriptions will be explained in the following sections.

As noticed, some features like rainfall and temperature, which are present in each dataset, don't go alongside the date. Indeed, both rainfall and temperature affect features like level, flow and hydrometry some time after it fell down. As we don't know how many days/weeks/months later rainfall affects these features, this is another aspect to keep into consideration when analyzing the dataset.

TABLE I
ARCHITECTURE OF MLP NEURON NETWORK (LAKE BILANCINO)

Layers	Dense	neuron = 8	activation = ReLu
	Dense	neuron = 8	activation = ReLu
	Dense	neuron = 8	activation = ReLu
	Output Dense	neuron = 2	activation = Linear
Loss function		MSE	
Batch size		64	
Epoch		200	
Optimizer		Adam	

IV. METHODOLOGY

A. Exploratory Data Analysis

There are a total of 6603 entires and 9 columns in the Lake Bilancino dataset, among which 6 columns are features and 2 columns are targets. For the River Arno dataset, there are 8217 entires and 17 columns in total. 15 columns are features and the last column is the target. Both dataset contain rows with NaN values, and this will be handled in the next step. The following plots display the analysis results on Lake Bilancino and River Arno.

B. Data Preprocessing

In this project, the NaN values are handled by deleting the entire row, and the data remains in chronological order.

Next, we split the data into training and testing sets.

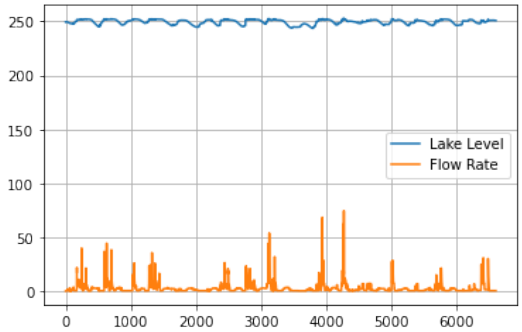
Furthermore, normalization has been performed on the training and the dataset, respectively, where every feature is normalized between 0 and 1.

C. Models

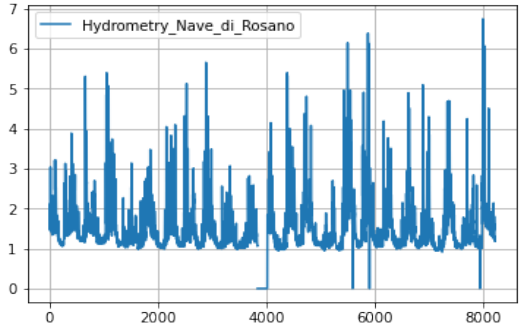
1) Multi-layer Perceptron (MLP) neural network: Multi-layer Perceptron neural network belongs to feedforward neural networks. MLP consists of multiple layers of neurons including an input layer, one or more hidden layers and an output layer. The weighted sum of neurons in each layer is past to the neurons in the next layer. Then it goes through their corresponding activation functions, which map the input to the output of this neuron. The architecture and settings for the MLP used is as displayed in Table I.

2) Long and Short-Term Memory (LSTM): In the LSTM network structure, the LSTM memory unit is located in the center, the input is known data, and the output is the predicted result. There are three gates in the memory unit, which can selectively add and filter the information passing through the structure. The three types of "gates" include: forget gates, input gates and output gates. The forget gate is used to control the historical information stored in the hidden layer node and at the previous moment. The input gate is used to control the input of the hidden layer node at the current moment. The output gate is used to control the output of the hidden layer node at the current moment.

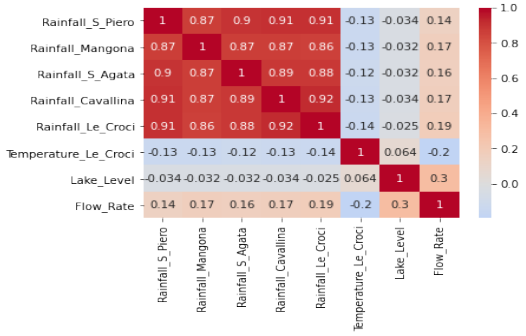
In the LSTM network, the input values need to be reshaped into 3-D as samples, where the time step is set 10, the input dim is set the same as the feature number, which depends on the specific sub-datasets. As described above, we need to



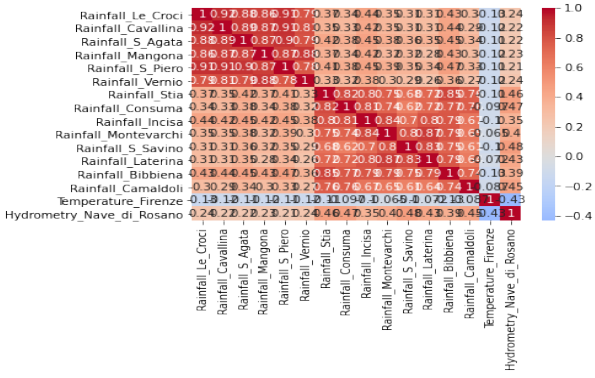
(a) Lake Level and Flow Rate (Lake Bilancino)



(b) Hydrometry (River Arno)



(c) Correlation Heatmap (Lake Bilancino)



(d) Correlation Heatmap (River Arno)

Fig. 1. EDA on Bilancino Lake and River Arno data

TABLE II
ARCHITECTURE OF FOR THE ARNO RIVER DATASET

Layers	LSTM	neuron = 128	activation = tanh
	Dense	neuron = 100	activation = ReLu
	Dense	neuron = 50	activation = ReLu
	Dense	neuron = 30	activation = ReLu
Loss function	MSE		
Batch size	64		
Time step	10		
Epoch	50		
Optimizer	Adam		

TABLE III
ARCHITECTURE OF FOR THE BILANCINO LAKE DATASET

Layers	LSTM	neuron = 32	activation = ReLu
	Dense	neuron = 30	activation = ReLu
Loss function	MSE		
Batch size	64		
Time step	10		
Epoch	50		
Optimizer	RMSprop		

split the training and testing sets. We chose the 0.2 as the test size for the sub-dataset Bilancine Lake and the Arno River for testing, and the left samplings for training. The input values have been reshaped into 3-D as samples, where the time step is set 10 and the input dim is set 8 and 16, respectively. In this way, the preprocessed data can be fed to LSTM to make predictions through historical and current data.

The network architecture for these two datasets are shown in Table II and Table III, some parameters are tuned by the hyperparameter-tuning processes, which will be discussed the next sub-section.

3) Support Vector Regression(SVR): A SVR model has a configuration that is somewhat similar to a three-layer ANN with a hidden layer with the same number and feature as the support vectors. To put it another way, the number of hidden nodes is the same as the number of support vectors. It's important to remember that in SVR, the whole model architecture is adaptively generated directly, while in ANN, only the weights are automatically generated. That is, while the adaptive SVR algorithm determines the number of support vectors, the number of hidden layers and hidden nodes for each layer must be artificially estimated in advance for ANN [2].

D. Hyperparameter Tuning

1) MLP: The hyperparameters were tuned using GridSearchCV in sklearn. Choices for optimizer are ['adam', 'SGD', 'rmsprop']. Batch size includes [64, 128], and the number of epochs include [200, 300, 400].

2) LSTM: We need to tune hyper-parameters for the LSTM model through the grid research approach. The epoch number is tuned from [50, 100, 200], Batch size is from [64, 128], and Neurons in the first layer is from [32, 64, 128]. The activation function is from ['relu','sigmod','tanh'], and the optimizer is from ['Adam','SGD','RMSprop'].

The tuned structure of the LSTM have already been shown in Table II and Table III.

V. RESULTS AND ANALYSIS

A. Metrics

After providing the appropriate methods, it is also critical to evaluate this kind of time-series problem. Therefore, accuracy measurements should be discussed. As our targeting problem can be classified into the regression field, the involved values include the predicted values and the ground truth, which can be represented by $\hat{\varphi}_i$ and φ_i for clarify.

The Mean absolute error (MAE) and the root mean square error (RMSE) can be used to evaluate the performance of the forecasting models. The definition are given by:

$$\mathrm{MAE} = |\hat{\varphi}_\mathrm{i} - \varphi_i| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(\hat{\varphi}_i - \varphi_i)^2} \tag{2}$$

Where $\hat{\varphi}_i$ is the forecast data, while φ_i is the ground truth.

B. Benchmarking

The applied MLP, LSTM, SVR and GRU models in this project will compared with each other, and the optimal one will be selected for the prediction purpose.

This subsection will be supplemented after all the models having been trained.

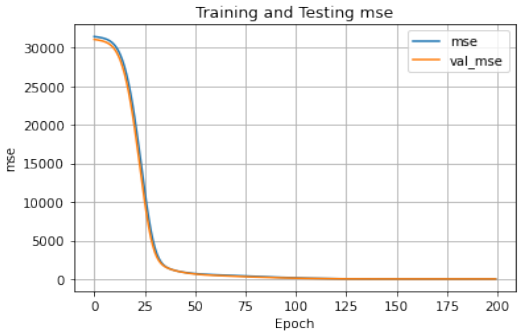
C. Results

1) MLP: As it is displayed in Fig. 2, the training and testing MSE decreases significantly as the number of epoch increases. The best MSE value is around 13.33 and the best validation MSE is around 12.54. As the difference between two errors are not large, mechanism for preventing overfitting problem is not adopted. Also, it is worth noted in the lake level and flow rate predictions plot that sudden increases or decreases are generally not predicted accurately. Events causing such abrupt changes can sometimes be totally random and therefore, they are very difficult to be predicted.

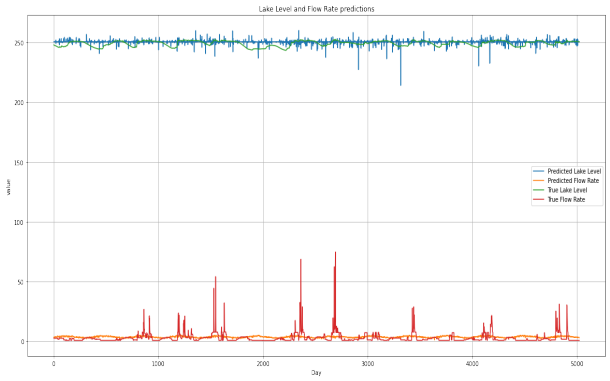
2) LSTM: Fig. 3 shows the virtualization of the experimental results, indicating the loss function values based on the training and the testing data during the training process. The X-axis is the epoch number, and the Y-axis shows the loss function values. Sometimes, the loss function of training data continuously decreases; however, the testing data does not.

To avoid overfitting, the early stopping mechanism is adopted, where the patience is set as 10. As the epoch number increases, if the test error is found to rise on the validation set in the continuous 10 epochs, the training has to stop. Then the values after stopping can be used as the final parameters of the network.

As known, there are three types of layers in the LSTM network, including the input layer, the hidden layer, and the output layer. The input to the LSTM network is the historical data. The output is the predicted data for a period of time.The number of memory cells is determined by the time step. The hidden layer number in Table II and Table III is one, but it can be easily extended to several layers according to the situation.



(a) MLP Training and Testing MSE (Lake Bilancino)



(b) Lake Level and Flow Rate predictions (Lake Bilancino)

Fig. 2. MLP results on Lake Bilancino data

3) SVR: Because the SVR has been cancelled once in this project. And the GRU model has not completed yet. The results of the SVR part are not shown in the progress report. It will display on the presentation and final report.

VI. Next Steps

A. MLP

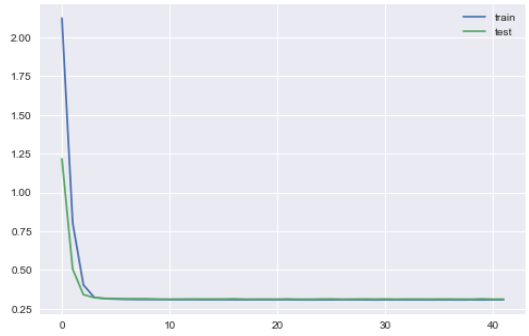
The goal for the future is to complete the predictions for Arno River, as the current model only predicts the lake level and flow rate of Lake Bilancino. Additionally, early stopping will be adopted to prevent overfitting problem and better improve the accuracy of the predictions.

B. LSTM

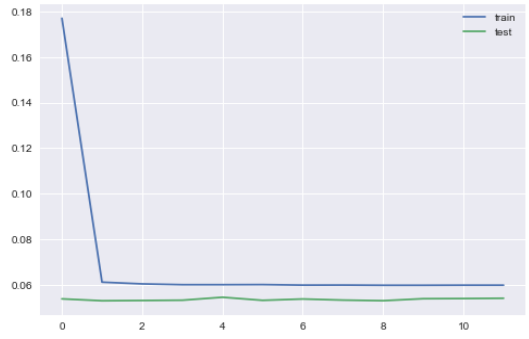
The current trained model for the prediction is more appropriate to the short-term data, for the long-term predication, more LSTM models with different layer numbers will be tested. Besides, some more hyperparemeters will be tuned as well. Finally, the LSTM-based model will be compared with other baseline models.

C. SVR

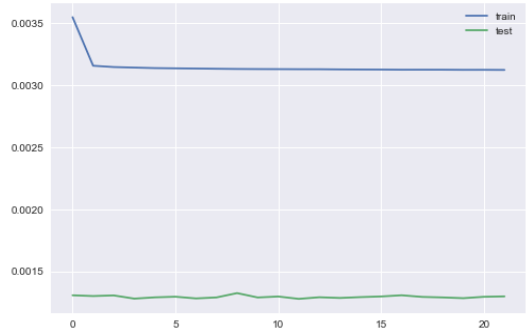
The current SVR part has redesign recently. There are more work will complete for the Arno River and Lake Bilancino. In addition, the goal of high accuracy will be achievement.



(a) hydrometry of the Arno River



(b) lake level of the Bilancino Lake



(c) flow rate of the Bilancino Lake

Fig. 3. Loss function values based on the training and the testing data.

References

[1] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[2] Ruas GIS, Bragatto TAC, Lamar MV, Aoki AR, de Rocco SM. Electrical energy demand prediction using artificial neural networks and support vector regression. In: 3rd International Symposium on Communications, Control and Signal Processing, ISCCSP 2008, IEEE; 2008

[3] "Acea Smart Water Analytics," Kaggle. [Online]. Available:*https://www.kaggle.com/c/acea-water-prediction/data*. [Accessed: 22-Mar-2021].