Data Prediction for Public Events in Professional Domains Based on Improved RNN- LSTM

Bonan Song, Chunxiao Fan, Yuexin Wu and Juanjuan Sun
School of Electronic Engineering, Beijing University of posts and telecommunications, Beijing, China
songguolan1@163.com, cxfan@bupt.edu.cn, wuyuexin@bupt.edu.cn, sunjj@bupt.edu.cn

Abstract. The traditional data services of prediction for emergency or non-periodic events usually cannot generate satisfying result or fulfill the correct prediction purpose. However, these events are influenced by external causes, which mean certain a priori information of these events generally can be collected through the Internet. This paper studied the above problems and proposed an improved model—LSTM (Long Short-term Memory) dynamic prediction and a priori information sequence generation model by combining RNN-LSTM and public events a priori information. In prediction tasks, the model is qualified for determining trends, and its accuracy also is validated. This model generates a better performance and prediction results than the previous one. Using a priori information can increase the accuracy of prediction; LSTM can better adapt to the changes of time sequence; LSTM can be widely applied to the same type of prediction tasks, and other prediction tasks related to time sequence.

1. Introduction
The growing amount of data greatly enriched the knowledge base of data; a lot of potential research values are embedded in these data. Domains such as electricity, transportation and telecommunications as deeply influenced by operational factors, and public events exactly include this key information. Therefore, information extracted from public events can be used as a priori knowledge, which contributes to the prediction of professional domains. In recent years, many scholars have conducted researches in the professional domains associated with time sequence, their research methods can be divided into: (1) methods based on ARIMA model or regression model [1]; (2) methods based on neural networks or deep neural network model of depth method [2].

Some of these researches combined with prior events, but most of them use the impact factor pre-set by industry experts, but few combined sources of information of public events with business events prediction of professional field of study. Although models based on ARIMA have the advantage of simple network and compatibility for time sequence related prediction tasks, still it is difficult for them to generate satisfying prediction because the models are unable to effectively analysis emergencies and non-regular periodic events or meet the assumptions of stationarity. Hence, these models cannot provide reliable results. Models based on neural network are relatively reliable on exploring non-linear factors, but most of these models used artificial neural networks. The introduction of deep learning is still in research; besides, its adaptability of time sequence is limited.

To solve the above mentioned problem, this paper proposes an LSTM algorithm based on deep neural networks and establishes a LSTM dynamic prediction a priori sequence generation model,
which use information obtained from public event sources as a priori information, and use the combination of prior information and data of professional domains as the input of LSTM model to generate sequence [3]. After the training of the model, a priori information of public events sources would be added to every step of the real-time dynamic forecast, together they will predict the result of next time step. This paper combines the LSTM algorithm and public domain data sources to train the model in the field of professional domains prediction tasks, experimental studies show that (1) in addition to traditional natural language processing, songwriter generation and so on, LSTM algorithm can also be adopted to support data prediction through sequence generation. And this model has a certain degree of versatility; (2) combined with a priori information of the public event sources, the model can correctly predict data event trends, which indicates that the model has a better accuracy; and (3) the improved LSTM algorithm can support the prediction of dynamic splicing new data.

2. Data prediction model for professional domains by combing priori information

2.1. Data of public domains events and data information sources of professional domains

Some of the emergencies and non-regular periodic events have been pre-arranged, such as arrangements for competition events, activity reports, broadcast programs, these types of information can be extracted prior to the event on the Internet. Therefore, these events are called prior public events. The reason that the predicted business data in professional domains is unable to fully reflect its regularity is because these events do not occur periodically and even trigger emergencies, but they still can use the a priori knowledge.

Based on these conditions, we divide the factors affecting business data of professional domains into two and extract the information of data sources individually. One is data sources of priori public events, whose data can be collected on the Internet, such as on sites with scheduled events and hot events on Weibo. Another is data sources of professional domain events, whose data are often managed by the data maintenance specialist, such as the operator's traffic flow information, traffic management information—vehicle flow and flow of people, public resources management information—electric power, and water conservation information.

Prior public events often have some of the features, such as:
- predictable: you can know event attributes such as size, duration, scope in advance;
- non-periodic: these events show limited periodically characters, because they happen according to specific arrangements, and sometimes unexpected events occur;
- obtainable: a large-scale of event data, and event attribute information is available on the Internet and easy to be collected.

2.2. Status quo of domestic and abroad researches

Currently most data prediction tasks of professional domains are confined in prediction tools or model algorithms such as probability statistics, gray model, Markov chain, ARIMA model and traditional neural network. And all of them use business data as the only source of data to predict. This paper proposes a model to improve the following aspects (1) most data prediction of professional domains was not combined with external a priori public information; (2) traditional business data prediction algorithms only applied limited time sequence and deep learning; (3) LSTM sequence generation algorithm is not sufficient to support dynamic a priori information splicing. Using LSTM-RNN (short and long term memory recurrent neural network) to build overall prediction model, this paper combines information from two event sources and uses attributes and characteristics of both sources as input to train the model. In the end of this paper, the prediction will be given.

2.3. LSTM dynamic prediction model for a priori sequence generation

First, both a priori public events and professional domain business prediction in this paper are related to time sequence. Time sequence prediction tasks face the following problems: (1) in the time sequence of the events are autocorrelation and non-stationary. It requires the model to adapt to the
autocorrelation and partial correlation of time sequence. At the same time, after the introduction of events influence, as it takes a long period of time to change, the attribute of time sequence is non-stationary. The non-stationarity should be set as a prerequisite for model design. (2) the business data is correlated in terms of time transfer, because they all use the information of previous sequence, which results in data memory. Therefore, the model needs to be able to memorize for a period of time and analyze the latest record of the event occurred and the data sequence changes of several previous events.

Second, most of the previous model predictions adopted closed internal data prediction, which is relatively simple and unable to analysis external data. Also, most of them use traditional neural network to predict the future business data distribution. These mathematical statistical models cannot locate the timing of public events. The stationary distribution assumption of ARIMA model does not meet the demand of this scenario. So this paper use neural network RNN-LSTM deep learning model to predict the result.

The working process of conventional LSTM sequence generation model is—establish model through input data training, submit predetermined time step data to the model through using it, thereby after a predetermined time steps, the model can use its parameters and the context provided by previous time steps to successively predict the entire generation sequence and produce the final prediction result. But such predictions model cannot analyze multiple data sources, therefore, they are unable to generate a dynamic prediction or reasonably display the data tendency of emergencies or non-periodic event of professional domains, or effectively improve the prediction accuracy of the model.

Thus, this paper uses the spatial and temporal information as the prerequisite and a priori knowledge of the new model. Deep learning neural network can be adapted to non-linear issues of time sequence. Also, based on the cyclic neural network RNN-LSTM, this model can both solve the problem on a long term memory of time sequence and improve the prediction accuracy by changing the network structure and adding context and a priori information to nerve cells of the network. All these factors will contribute to a more accurate prediction.

3. Introduction of the Technology

3.1. Recurrent Neural Networks

Recursive neural network (shown in Fig. 1) is an artificial neural network for processing timing sequences. It is a cyclic neural network that can be used in tasks such as speech processing, picture description, text generation, text translation, sequence generation and prediction, etc. Because it adopts an error calculation method called Back-Propagation Through Time (BPTT), in which the traditional RNN cannot solve the data association problems that depends on long time. This results in disappearance or the explosion of the gradient. In order to solve the problem, experts including Hochreiter proposed an improved LSTM (long short term memory) in 1997 to resolve the gradient disappearance [4].

As is shown in Figure 2, LSTM proposes a gate mechanism to determine the accumulation of information. It is mainly dealing with problems of sequence of time. Its availability and accuracy in domains including natural language processing and voice processing are proven [5].
LSTM is composed by input gate, forget gate and output gate. As the basic logic units, these gates match the data through weight and the activation function, and control the input, output and forget part of the entire network via the gate. Each neural network module would transfer the message in a cell to the next module. In the following formula, $i_t$, $C_t$, $f_t$ and $O_t$ represent the input gate, the state of the cell, forget gate and output gate corresponding to $t$. LSTM uses the state of cells to record information transferred in neurons. It runs across the neural network with linear interaction. The error function of selective memory feedback is adopted to adjust network weights, forming the predictive model. This allows the long-term message to be passed along the time in multiple recursive neural networks.

3.2. Structure and Principle of LSTM

Cell state

(1) Forget gate: The forget gate is composed by sigmoid activation function, weight matrix $W_f$ and bias matrix $b_f$. $x_t$ refers to the input information at the current time step. $h_{t-1}$ refers to the corresponding output of the time step.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

(1)

(2) Input gate update information: Input gate controls the influx into neurons. $i_t$ refers to the information to be updated of the time step generated by the sigmoid gate. $\tilde{C}_t$ refers to the retained information of the time step calculated through tanh.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

(2) (3)

(3) Updated state of the cell: The equation calculates the cell status. $C_t$ refers to the state of the cell of the present time step. $C_{t-1}$ refers to the one of the previous time step.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

(4)

(4) Output gate: A control gate output is the output of neuron portion. After the layer is obtained using the sigmoid cell state requires the output portion, is through Layer determines which cells need to be output state.

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = O_t \cdot \tanh(C_t)$$

(5) (6)

LSTM sequence generation can fully demonstrates the regular change of certain data. It functions the best when the time is fixed and the change is regular. When processing natural language, for example, the sequence generation of the text, after a large amount of training, the model can memorize the language style and language features. Given certain input, it can predict changes in certain time period, and can memorize the features of the provided text, record them in the cell status, observe the features, and provide the corresponding prediction.

4. Data prediction of public events based on improved RNN-LSTM

4.1. Experiment Design

(1) Using information source of public events

Information of public events is collectable. One can easily get its properties, such as duration (time), scale, and scope of activities and so on. In this study, relative attributes are duration and weight of the public event influence. These properties are quantified, and combined with influence weights provided by experts. Together they serve as the weight of the public event. Weights are categorized by its volume in specific area, such as concert’s weight is heavier than the one of sport events. Information of weights of common events shall be gathered, such as small exhibition venue, events and games. When there is no public event, fixed constants are set, such as zero in usual time, or the peak volume. (The follow-ups of public events are not discussed in this paper).

(2) LSTM sequence generation model

This paper intends to use the predictive function of LSTM sequence generation, and incorporate the source data into data from specific areas in public domain. By designing the training and validation format, data extracted from public events are put into LSTM training mode, improve the verified data
source, dynamically assemble public data sources at each time step, so as to provide prediction when the data of the initial time step is given and data of each following time step is incorporated.

4.2. Dataset

4.2.1. The combination of data of public domain data and professional domain. In this study, data sets of the volume of base station round Beijing Workers’ Stadium are incorporated, with the 4G data volume of Beijing Workers Stadium as the main body of the predictive model. The data sets are divided into a training set and test set.

With specific attributes of a specific event, we can predict the volume of a specific business, such as telephone traffic, electricity flow, traffic flow, visitor flow and so on. These data of specific areas would be influenced by non-periodic events or incidents. Therefore, it requires information of events in public domain.

4.2.2. Pre-processing data

(1) Public event correlation model: Collect attributes in public events, and assign weights to public events according to their scales. Put the corresponding events in the time step in accordance with the event’s time information and form new training data set.

(2) Normalization: Process date to (-1, 1)

(3) Sliding window: Change to the mode of sliding window. The time step is the variation. Pile the training data and test data in the form of sliding windows to achieve a better fit results.

\[
x_1, x_2, \ldots, x_n = x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{2n} \quad (n = \text{timestep})
\]

\[
y_1, y_2, \ldots, y_n = y_1, y_2, \ldots, y_n, y_{n+1}, \ldots, y_{2n} \quad (n = \text{timestep})
\]

4.3. Establish improved LSTM model

This article is based on Tensor flow framework of Python and RNN-LSTM to build the model. It uses sequence to sequence LSTM the network to establish an unidirectional RNN network with LSTM of 4 base unit layers.

(1) Introduction of training mode

When establishing four-layer LSTM model, three to four hidden layer can make it faster to converge, effectively reduce the number of cell units in the hidden layer, and shorten the training time. More hidden layers are more likely to cause the model to overfit.

This study uses four-layer LSTM network as training model, and input both \( x_e \), data of public events, and \( x_p \), data of specific area as the input \( x(x_e, x_p) \). The design of four-layer LSTM model is as shown in Figure 3.

\[ h_t \] is the hidden layer node unit of \( h_1^t \) to \( h_4^t \) is four hidden layers. \( t \) is the time step.

A hyperparametric \( n \) is required before training. \( n \) refers to certain time steps. The sequence observes \( n \) steps of input and calculates the true value of the output. The error between the prediction at the end of \( n \) steps and the previous batch is calculated. The error is passed back along the timeline and simulate fitting. During training, \( n \) steps of input would enter \( t \) time steps. Data of each time step \( t \) is as shown in equation:
\[ y_{\text{time step}} = y_{11}, y_{12}, \ldots, y_{1n} \] (9)

In each batch, each time step \( t \) would produce a predictive sequence value \( y_t \). Together with the expected output value of the sliding window \( \text{pred} \), we can calculate out cost. The model adopts squared error function to calculate error return. \( \text{cost} \) is the minimized target of the calculation, as shown in the following equation:

\[ \text{cost} = (\text{pred}_t - y_{1t})^2 \] (10)

(2) Improved forecasting series model way

As is shown in Figure 4, the sequence generation method uses the result of a time as the input of following time step. Dynamically input the weight of the public event of next time step as the input value the next prediction.

\[ x_t = (y_{t-1}, x_{e_{t-1}}) \] (11)

(3) Testing accuracy:

Testing data: Testing data requires \( p_1 \ldots p_n \), data of a specific area of a certain time step \( n \), as well as data of events in public domain of the correspondent time. After the iteration of 70 steps, it generates the prediction data of \( 70-n \) step. As described above, output \( y \) of each time step would be the input of the next time step. Together with the dynamic incorporation of data from public events, they will form a new input to predict. And perform the ultimate verification of the model’s accuracy of result and tendency.

\[ y_t = \text{previous}(\text{tuple}(x_t, x_{e_{t-1}}), y_{t-1}) \] (12)

5. Experimental results and analysis

This is experiment is based on Tensor Flow depth learning framework and improved algorithm of LSTM in python. Tensor Flow is an artificial intelligence system google released. Compared to CAFFE, it is more adapted to the cycle of neural networks and deep learning scenario of LSTM. Compared to Torch and THEANO etc., it has better adaptability and is more easily to be understood.

5.1. Model training and evaluation

(1) Hyper-parameters

Required experimental initial parameters are: learning rate, number of hidden layers, the number of hidden layer nodes, gradient descent clipping ratio, the iteration number, batch number etc. The experiment maintains a single variable to calculate the error gradient of the sequence adopting forward transmission method of BPTT, and to update the weight values. The experiment has four unit layers, LSTM hidden layers and input layers connected to the front (back) of the hidden layer. When there is no significant reduction of error, the iteration terminates, so that the model converged.

(2) Evaluation

To evaluate the model and the prediction of the model, this paper will apply parts of the public events information, and uses a ratio of the predicted value and the real values to showcase the accuracy of the prediction.

The tendency of the volume of the specific area is illustrated in line chart, in order to show whether the prediction of the tendency is correct when the impact of public events is intense.

5.2. Training results

(1) Business trend data to predict professional domains

As is shown in Figure 5, the predictive model functions well when predicting tendency. It can capture small fluctuations of data of small events, as well as the increasing regions of the change of the data when public events happen. It can also simulate rising/falling speed of the data when it is supposed to change. And when there is no impact from public events, the prediction is very accurate.

Meanwhile, the prediction of peak value under the influence of public events is not accurate enough. But it can provide solid prediction of the information of when the value is getting to the peak. In addition, the prediction of falling region under the influence of public events is not accurate enough.
Due to the error in decreasing region would be passed down with time, the prediction of data after the public event is damaged to some extent.

(2) Model’s relative accuracy
As described above, the location of large public events is step 38-50 in the figure. As shown in Figure 6, error is small when there is no public event happening, and the prediction is more accurate. When the prediction for large public event falls in the declining region, the accuracy is not good. In this region, the error is relatively bigger. Since the data has a substantial magnitude, normalization will result in amplification of small error. But the capture of public events and incidents and the prediction in rising region are relatively accurate. Error ratio can be controlled at about 12%. The model may occasionally make a larger error, but it does not affect the subsequent prediction.

Figure 7 shows the error ratio of large-scale public events (ie 38-50 steps). It shows that the prediction after the event is less accurate than the one of the rising region. Accuracy is more guaranteed when the event is rising.

5.3. Analysis of training results
(1) The number of cells in hidden layers: the size of cells in hidden layer is 140. It proves that the higher number of cells in hidden layers can perform better fit with the data peak.
(2) LSTM layers: the experiment adopts RNNCELL network with 4 layers. The more layers are, the faster convergence of the model is. When there are four layers, the gradient explosion is less common.
(3) Dynamic learning rate: the method does not drop when the anti-conditions, the application rate of dynamic learning occurs in the model accuracy, a small decrease learning rate, this model was observed to a maximum consumption current batch error before two batches of training average error consumption, decreased learning rate of 0.8 times the current, in order to improve the fitting accuracy.
(4) Learning rate: The model’s learning rate is 0.007. The higher the learning rate is, the better the fit to the peak value would be; the lower the learning rate is, the better the fit to the valley value would be.
(5) Time step: The prediction is conducted every 24 hours in this study. Since the unit is hour, we adopt 24 hours a day as a period of time, which is a relatively short time for fitting.
(6) Batch: The number of batch is 475. The experiment shows that in each batch of training data, a large change in specific area shall be included to produce more accurate model. The outcome is more ideal when extreme values are included.

Because of the need of predetermined setting of time steps when generating a sequence, which means the prediction can only be performed after these time steps, when the dataset is in sufficient, the usage of sliding windows can largely improve the fitting as it returns the error after the prediction of each time.

Conclusion
The model can correctly predict the tendency of change after public events causes data changes. It can also capture fluctuations caused by small events. The accuracy of the model is consistent. The prediction of the time steps at the peak is less accurate than common, but the prediction of data’s tendency is correct. The overall performance of the model in a long sequence generation tasks is in line with expectations. LSTM model can provide relatively accurate prediction by sequence generation when proper information is given.

Error Analysis: Due to the sequence generation method, the error in the previous steps would be accumulated, and passed down to the later prediction. The error is bigger when large-scale public events occur. However, it does not significantly sabotage the prediction.

Improvement: More data of the volume can be input to improve accuracy. Input more influence of public events into the training dataset in order to improve the prediction of peak value and the model of influence.

References
[1] Erasmo Cadenas, Wilfrido Rivera, 2010, Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN model. Renewable Energy
[2] Wu JiaoJiao, 2015, The Research of Gas Concentration Prediction Based on Space-time Neural Network Model, China University of Mining and Technology
[3] Alex Graves, 2013, Generating Sequences with Recurrent Neural Networks, computer science
[4] Sepp Hochreiter, 1997, Long Short-Term Memory, Neural computation
[5] Hu XinChen, 2015, Research on semantic relation classification based on LSTM, Harbin Institute of Technology
[6] Kai Sheng Tai, 2015, Richard Socher, Christopher D. Manning, Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, computer science