Prediction of Crop Pests and Diseases in Cotton by Long Short Term Memory Network

Chapter · July 2018

DOI: 10.1007/978-3-319-95933-7_2

4 authors, including:

Peng Chen (陈鹏)
Anhui University
137 PUBLICATIONS **1,709** CITATIONS
SEE PROFILE

Bing Wang
Anhui University of Technology
163 PUBLICATIONS **2,268** CITATIONS
SEE PROFILE

Prediction of Crop Pests and Diseases in Cotton by Long Short Term Memory Network

Qingxin Xiao[1], Weilu Li[1], Peng Chen[1(✉)], and Bing Wang[2(✉)]

[1] Institute of Physical Science and Information Technology, Anhui University, Hefei 230601, Anhui, China
pchen.ustcl0@yahoo.com
[2] School of Electrical and Information Engineering, Anhui University of Technology, Ma'anshan 243032, Anhui, China

Abstract. This paper aims to predict the occurrence of pests and diseases for cotton based on long short term memory (LSTM) network. First, the problem of occurrence of pests and diseases was formulated as time series prediction. Then LSTM was adopted to solve the problem. LSTM is a special kind of recurrent neutral network (RNN), which introduces gate mechanism to prevent the vanished or exploding gradient problem. It has been shown good performance in solving time series problem and can handle the long-term dependency problem, as mentioned in many literatures. The experimental results showed that LSTM performed good on the prediction of occurrence of pests and diseases in cotton fields, and yielded an Area Under the Curve (AUC) of 0.97. The paper further verified that the weather factors indeed have strong impact on the occurrence of pests and diseases, and the LSTM network has great advantage on solving the long-term dependency problem.

Keywords: Long short term memory · Weather factors · Recurrent neural network · Occurrence of pests and diseases

1 Introduction

Recently, the occurrence frequency of regional cotton pests and diseases has increased rapidly, causing huge losses in agricultural production. There are many factors to affect its growth, of which the most significant one is abnormal climate change, which resulted in the continuous evolution of pests and further made them adaptive to the environment. All of that seriously influenced the yield and quality and made it very difficult to control the pests and diseases [1]. Many methods have been developed to control pest occurrence. One type was based on biochemical perspectives to suppress the occurrence of pests and diseases, i.e., pesticide screening [2], biological control [3]. The other type was based on historical data and tried to predict future occurrence trend of pests [4].

In recent years, deep learning has been widely used in many fields [5–10]. Long Short Term Memory (LSTM) is a deep learning model. It is a special kind of recurrent neutral network (RNN), which introduces gate mechanism into vanilla RNN to prevent

D.-S. Huang et al. (Eds.): ICIC 2018, LNCS 10955, pp. 11–16, 2018.
https://doi.org/10.1007/978-3-319-95933-7_2

the vanished or exploding gradient problem. LSTM has achieved good results in different fields. Li *et al.* adopted an LSTM auto-encoder with generating coherent text units from neural models to preserve and reconstruct multi-sentence paragraphs [8]. Gao *et al.* presented an mQA model, which contained a LSTM and a Convolutional Neural Network (CNN), to answer questions about the content of an image [9]. Theis and Bethge introduced a recurrent image model based on multi-dimensional LSTM units, which are particularly suited for image modeling [10].

In this paper, we propose a LSTM network based method to predict the occurrence of pests and diseases of cotton, with the use of weather factors. Results showed that our LSTM based model outperformed other traditional prediction models.

2 Methodology

2.1 Material and Problem Formulation

To investigate the impact of weather factors on the occurrences of pests and diseases, the datasets from Crop Pest Decision Support System (http://www.crida.in:8080/naip/AccessData.jsp) were used, which recorded cotton documents weekly (15, 375) for 10 insect pests and diseases in cotton along with corresponding weather conditions across 6 important locations in India. The weather features consist of Maximum Temperature (MaxT(°C)), Minimum Temperature (MinT(°C)), Relative Humidity in the morning (RH1(%)), Relative Humidity in the evening (RH2(%)), Rainfall (RF(mm)), Wind Speed (WS(kmph)), Sunshine Hour (SSH(hrs)) and Evaporation (EVP(mm)). Our aim is to predict the occurrences of pests and diseases under different weather conditions. Bollworm is the main target of biological control. So, bollworm records were used to build weather-pest forecasting model. Suppose X be the vector set of weather-pest records, $Y = \{0, 1\}$, from one single area along the whole recorded time, which is a time series set. The prediction problem can be then converted into predicting the occurrence ($Y_i = 0$) or non-occurrence ($Yi = 1$) of pests and diseases based on the feature vector X_i, $i = 1 \ldots N$, where N is the number of feature vectors. A LSTM based model was designed to capture the relationship of data (X_i, Y_i), $i = 1 \ldots N$, to predict future occurrence of pests and diseases under the weather features.

2.2 Long Short Term Memory

Like most RNNs, LSTM contains a memory function that can handle time series problems, while unlike traditional RNNs, LSTM is well-suited for long-term dependency problems because it solves the problem of gradient vanish and gradient explosion. There are three doors in LSTM. The input gate decides the input x_i entering into the current cell, the forget gate decides if and how much information be forgotten for the previous memory, and the output one controls the information outputting from the current cell. The gating operations ultimately determine which information is forgot and which information is entered into the neural network as useful information. For the weather-pest forecasting issue, it processes a series of temporal dependency inputs x_t at

time t and the hidden vector h_{t-1} from the last time step then get the predicted h_t. The basic structure of LSTM cells can be seen in Fig. 1 and the related formulas are shown as below.

$$
\begin{aligned}
i_t &= \sigma\left(W^i \cdot [h_{t-1}, x_t] + b^i\right) \\
f_t &= \sigma\left(W^f \cdot [h_{t-1}, x_t] + b^f\right) \\
o_t &= \sigma(W^o \cdot [h_{t-1}, x_t] + b^o) \\
\mathrm{c}_t &= \tanh(W^c \cdot [h_{t-1}, x_t] + b^c) \\
C_t &= f_t \cdot C_{t-1} + i_t \cdot c_t \\
\mathrm{h}_t &= o_t \cdot \tanh(C_t)
\end{aligned}
\tag{1}
$$

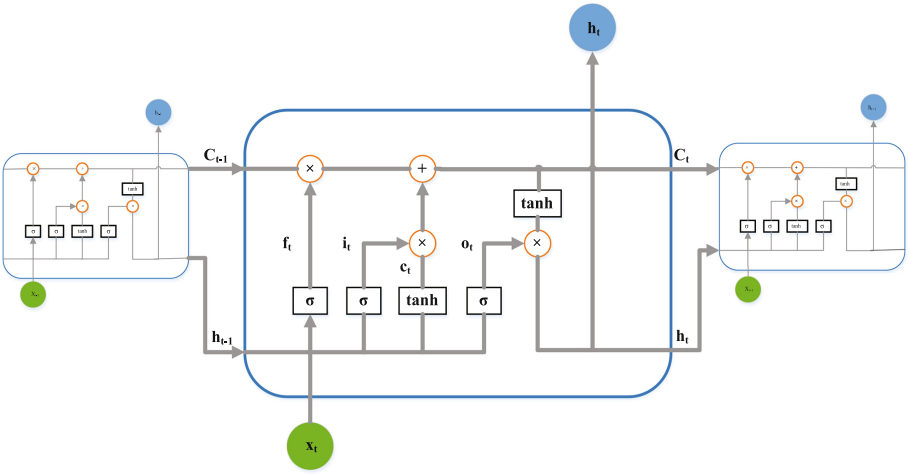

Fig. 1. Structure of LSTM cells.

where σ is the sigmoid function; *tanh* (*) is a nonlinear activation function; W is the recurrent weight matrix; b is the corresponding bias vector; i, f and o are the outputs of the input, forget, and output gates, respectively; and C and h are the memory vector and out vector of the cell, respectively.

According to the previous work [11], the output, (h_t, C_t), of a cell can be represented as a whole function *LSTM* (*):

$$
(h_t, C_t) = LSTM([h_{t-1}, x_t], C_{t-1}, W) \tag{2}
$$

where W concatenates the four weight matrices W^i, W^f, W^o and W^c.

2.3 Architecture of the Used LSTM Network

The prediction problem is converted into a time series problem, which uses the past weather-pest records to identify whether pests and diseases will occur in the future. The first thing that should be determined for the LSTM network is how long the historical observations is used for the prediction. Of course the longer the historical data is, the better the prediction will be, however the LSTM requires more computation. Here 'timesteps' is set as 4, i.e., four samples of weather-pest data are input together into the LSTM. Three other important parameters for the whole structure of the network should also be determined: the number of layers for LSTM layer l_r, the full-connected layer lfc and the corresponding number of hidden units denoted by *units_r*. In addition, some critical parameters have to be determined, i.e., the Adam optimization method [12] is adopted, and the learning rate and dropout are set as 0.001 and 0.1, respectively.

3 Experiment and Results

3.1 Determination of Parameters

Five top size datasets, denoted as p1, p2, p3, p4 and p5, were selected to train the LSTM network and determine the parameters. In this work, Accuracy (*ACC*), Area Under the Curve (*AUC*) and *F1-score* are used to measure the effectiveness of prediction methods. First, supposed l_r, l_{fc} and *units_fc* be 1, and set a proper value of *units_r* from {4, 5, 6, 7}. Table 1 shows the predictions on five datasets with different values of *units_rs*. The boldface items in the table represent the best performance. It can be seen from the table that the best performance occurs when *units_r* = 5 on three datasets p1, p2 and p4. Although the model performs not good enough on datasets p3 and p5, the performance differences are not obvious. Therefore, *units_r* was set as 5 in this work. Then, the determination of the proper value for l_r and l_{fc} from {1, 2, 3} is in the same way. The experimental results showed that the best performance occurs when l_r = 1 and l_{fc} = 2. The reason may be due to the increasing number of weights with

Table 1. Predictions on five datasets in terms of *units_rs*.

Units_r	Metrics	P1	P2	P3	P4	P5
4	ACC	0.9241	0.8973	0.9111	0.9017	0.8742
	AUC	0.9712	0.9532	0.9687	0.9578	0.9465
	F1-score	0.8857	0.8258	0.8316	0.8737	0.7804
5	ACC	**0.9329**	**0.9169**	0.9176	**0.9136**	0.8903
	AUC	**0.9764**	**0.9674**	0.9663	**0.9704**	**0.9715**
	F1-score	**0.8949**	**0.8555**	0.8580	**0.8955**	0.7903
6	ACC	0.9281	0.9063	0.9098	0.8949	0.8968
	AUC	0.9737	0.9643	0.9529	0.9628	0.9649
	F1-score	0.8896	0.8450	0.8420	0.8680	**0.8234**
7	ACC	0.9276	0.9013	**0.9255**	0.9000	**0.9032**
	AUC	0.9710	0.9557	**0.9717**	0.9551	0.9636
	F1-score	0.8870	0.8205	**0.8584**	0.8763	0.8104

increasing network layers, which results in lacking of insufficient data to train LSTM with large amount of weights. So we set $l_r = 1$ and $l_{fc} = 2$ to build the basic framework of the LSTM. We hope that the model has good generalization and could be applied in different cotton pests and diseases, so other pests and diseases records, such as jassid, whitely, and leaf blight, are input into the model to show the prediction power. The performance comparison on different kinds of datasets with LSTM network is listed in Table 2. Our model not only performs well in pests prediction, but also in disease prediction.

Table 2. Predictions on different kinds of pests and diseases with LSTM network.

Metrics	Bollworm	Whitefly	Jassid	Leaf blight
ACC	0.9207	0.9244	0.9354	0.9557
AUC	0.9719	0.9687	0.9776	0.9868
F1-score	0.8749	0.9243	0.9161	0.9204

3.2 Prediction Comparison with Other Methods

The bollworm dataset "p1" was adopted to implement the prediction comparison of our proposed method with other classical machine learning methods KNN [13], SVC [14] and Random Forest [15]. Their parameters were set as follows. For LSTM network, the parameters of *units_rs*, l_r and l_{fc} were set as 5, 1 and 2, respectively; for KNN, *n_neighbors* was set as 3; for SVC, LinearSVC was adopted and C was set as 10; for Random Forest, *n_estimators* was set as 100. Table 3 lists the prediction results. It can be seen from the table that LSTM network achieved the best prediction performance.

Table 3. Performance comparison on dataset "p1" with different methods.

Methods	ACC	AUC	F1-score
KNN	0.8426	0.8246	0.7365
SVC	0.7400	0.6353	0.4609
Random Forest	0.8563	0.8197	0.7481
LSTM network	**0.9174**	**0.9690**	**0.8603**

4 Conclusion

In this paper, we convert the problem of cotton pests occurrence into a time series classification problem. This is the first time, to our knowledge, to use LSTM to solve this prediction problem. The model could predict the occurrence of cotton pests and diseases according to weather conditions in the future, so that people can take real time precaution and reduce crop economic losses. Then, we also investigated the model on different types of cotton records, and achieved good predictions. In addition, some traditional machine learning methods were implemented to show the prediction

comparison with LSTM model. Results showed that LSTM has certain advantages in processing time-dependent problem. However, this paper only addresses the issue of the occurrence of cotton pests and diseases and predicts the occurrence with respect of weather factors. So, in the future, we will construct model to predict the hazard level of pests and diseases, so that prediction results are more responsive to data, making it easier for people to develop detailed pest control strategies.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Nos. 61672035, 61300058 and 61472282).

References

1. Wu, K.M., Lu, Y.H., Wang, Z.Y.: Advance in integrated pest management of crops in China. Chin. Bull. Entomol. **46**(6), 831–836 (2009)
2. Luo, J.Y., Zhang, S., Ren, X.L., et al.: Research progress of cotton insect pests in China in recent ten years. Cotton Sci. **29**, 100–112 (2017)
3. Satnam, S., Mridula, G., Suneet, P., et al.: Selection of housekeeping genes and demonstration of RNAi in cotton leafhopper. Amrasca biguttula biguttula. **13**(1), e0191116 (2018)
4. Zhang, W.Y., Jing, T.Z., Yan, S.C.: Studies on prediction models of Dendrolimus superans occurrence area based on machine learning. J. Beijing For. Univ. **39**(1), 85–93 (2017)
5. Huang, D.S.: Systematic Theory of Neural Networks for Pattern Recognition. Publishing House of Electronic Industry of China, May 1996
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
7. Graves, A.: Generating sequences with recurrent neutral networks. Comput. Sci. (2013)
8. Li, J.W., Luong, M.T., Dan, J.: A hierarchical neural autoencoder for paragraphs and documents. In: ACL 2015, v2, 6 June 2015
9. Gao, H.Y., Mao, J.H., Zhou, J., et al.: Are you talking to a machine? Dataset and methods for multilingual image question answering. arXiv:1505.05612 (2015)
10. Theis, L., Bethge, M.: Generative image modeling using spatial LSTMs. arXiv:1506.03478 (2015)
11. Kalchbrenner, N., Danihelka, I., Graves, A.: Grid long short-term memory. Comput. Sci. (2015)
12. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. Computer Science. arXiv:1412.6980. (2015)
13. Coomans, D., Massart, D.L.: Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules. Anal. Chim. Acta **136**, 15–27 (1982)
14. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
15. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)