*Journal of Computing and Electronic Information Management*

ISSN: 2413-1660

# Lightning Prediction Model Based on Spatial Density Clustering and CNN-LSTM

Hongbin Huang [1], Keyuan Tang[1], Lecai Cai [2]

[1]Sichuan University of Science & Engineering, Sichuan 644000, China

[2]Yibin University, Yibin 644000, China

**Abstract:** Lightning has a great impact on the stable operation of the power system. Based on the data of the lightning monitoring system, the tracking and prediction of lightning cloud clusters are realized. First, use the grid method to count the lightning data; then, use spatial density clustering to cluster the lightning points into lightning cloud clusters, and determine the center of gravity of the thunder cloud cluster according to the clustering results; take the time and geographic location of the thundercloud as The input is integrated into the same dimension; finally, the CNN-LSTM algorithm is used to dynamically predict the movement of the lightning center of gravity, so as to realize the tracking and prediction of thunderclouds. Taking the monitoring system data of a power supply bureau as an example, the simulation analysis was carried out in matlab2019, and the results showed that the prediction results of the algorithm compared with SVM and LSTM were more accurate and more practical.

**Keywords:** Lightning forecast, Spatial density clustering, Convolutional Neural Network, Long Short-Term Memory Network.

## 1. Introduction

Lightning has a huge impact on the normal and stable operation of the power system. In recent years, with the continuous development of China's power system, the increasing scale and complexity of the power grid, and the increasing number of extreme weather, the reliability of the power system increasingly requires stronger guarantees. Accurately tracking and predicting the direction and coverage of thunderclouds is one of the key measures for active lightning protection of the power system. [1] Therefore, it is necessary to better grasp the law of changes in the generation of lightning, and guide all power companies to construct lightning protection projects and formulate lightning protection plans accordingly. , Both have a positive meaning. Nowadays, all provincial and important municipal power companies have established the Lightning Location System (LLS), which has obtained and accumulated a large amount of data. With the introduction of data mining technology, a large number of lightning data are used by experts and scholars for standardized processing and statistical calculations, and conduct regular analysis and forecasting research on lightning.

Among them, the problem of how to more accurately describe and predict the direction and coverage of thunderstorm clouds can provide new solutions to problems such as the location and probability of lightning strikes and the calculation of lightning current magnitude. Literature [2] uses the data of lightning and radar to adopt a new hybrid charged particle tracking and monitoring method, which is based on the pattern recognition method and the principle of convective precipitation, and calculates the distribution of heavy precipitation and the range of active thunderclouds separately. Then merge. Literature [3] identifies the boundaries of cloud clusters and adds topological processing to establish the relationship between the life sequence and genealogy of cloud clusters, and establish the time sequence of each cloud cluster at the same time. The above methods can track and predict more accurately, but generally combine radar data, precipitation and other meteorological data, and have strong requirements for the completeness of the data, and it is difficult to accurately confirm the correlation between the elements; and the methods are mostly studied. The changes in thunderclouds have not paid enough attention to the actual data and fluctuations of thunderstorms.

In view of the above situation, the author uses the LLS of a certain area to collect data, and takes the number of lightning strikes in the area as the research object. The observation area is divided into equal time periods, and then the area is gridded and the number of lightning strikes is clustered., Transform the prediction of the movement law of thundercloud into the prediction of its center of gravity and shape change, and use the obtained state value as the input object, and use the CNN-LSTM algorithm to predict the further change of thundercloud. At the same time, the comparison with other existing methods confirms the feasibility and practicability of this method.

## 2. Spatial density clustering

This paper first defines the time and space scale of thundercloud prediction; the speed of thundercloud is about 1 km/min. Literature [4] found based on statistical data that the range of thunderclouds no longer continues to expand after a continuous thunderstorm for 5 minutes. After that, the duration of thunderclouds generally ranges from tens of minutes to several hours, which varies greatly from region to region. Based on this, this article takes 5 min as the standard, and assumes that the thundercloud remains stationary within 5 min, and moves in jumps between different time intervals.

### 2.1 Lightning grid statistics lightning data

Using grids to study the statistical model of lightning strikes was first proposed by Wuhan High Pressure[5], which divides the geographic coordinates of the target area into grids at equal intervals, and calculates the density of ground lightning strikes by studying the regularity of the LLS data in each grid . Since the same longitude difference can represent the same distance, but different latitudes represent different distances due to the different radius of the spherical surface, the specific calculation：

$$\begin{cases} x_j = \Delta\alpha \cdot \pi R / 180 \\ x_w = \Delta w \cdot \pi R \cos w / 180 \end{cases} \quad (1)$$

Where, $x_j$ , $x_w$ represents the side length of the grid under longitude and latitude respectively; R is the radius of the earth;, $\Delta\alpha$ , $\Delta w$ are the difference between longitude and latitude respectively; $\cos w$ is the latitude of the grid, when the latitude is small, $\cos w$ changes little with $w$, so press The grid divided by latitude and longitude can also be approximated as a region of equal area

The division of different grid spacings affects the density of lightning points in the grid, and thus also affects the calculation amount and prediction results. The scale of a single thundercloud is mostly from several kilometers to more than ten kilometers. Combined with the inherent error that may exist in LLS (the accuracy of the core area is about 0.5 km, and the edge may reach 1 to 3 km), the grid spacing is selected to be 0.01° (about For 1 km)

According to the weather and distance scale, this paper defines the number of lightning strikes within 5 min within a 0.01°×0.01° grid area as the lightning strike density, and clustering calculations of cloud clusters are performed based on this.

### 2.2 Thundercloud clustering and center of gravity determination based on spatial density

In two-dimensional space, the clustering algorithm that is used more today is the clustering algorithm based on spatial density [6] (Density-Based Spatial Clustering of Application with Noise, DBSCAN), which defines clustering as based on "density connection The central idea of this method: In the same cluster, with each element as the center in the neighborhood of a given radius (Eps), the number of data contained in the element is greater than a certain threshold (MinPts), that is, any A cluster set can be uniquely determined by any element. Among them, the elements (grids) with density greater than the threshold in the space can be called core points. definition:

The density is directly accessible. If there are elements $m$ and $n$, where $m$ is the core point and $n$ is in the neighborhood of m, that is $n \in D_{eps}(m)$ . The density of m and n is directly reachable.

The density is up to. There is $\left(x_1, x_2 \cdots x_k\right)$, there is $n = x_1, m = x_n$, $x_i$ and $x_{i+1}$ density are directly reachable. It is said that the density of $m$ and $n$ can be reached. After the cluster cloud is determined, the center of gravity position of the cloud can be obtained. Suppose the center of gravity of a single cloud cluster is (X, Y), the center coordinate of the i-th grid in the cloud cluster is, and the lightning density is, then the center of gravity of the cloud cluster is

$$\begin{cases} X = \sum_{i=1}^{n} \delta_i x_i \Big/ \sum_{i=1}^{n} \delta_i \\ Y = \sum_{i=1}^{n} \delta_i y_i \Big/ \sum_{i=1}^{n} \delta_i \end{cases} \quad (2)$$

By calculating the position and offset of the center of gravity of similarly located clouds in multiple continuous time spans, the trajectory and law of its movement can be determined, and then by counting the changes in the formation and movement of multiple thunderclouds, it can be dynamically tracked and forecast

## 3. CNN-LSTM prediction model

### 3.1 Convolutional Neural Network

Convolutional Neural Network (CNN) is a multi-layer neural network, which is good at processing images, especially large image related machine learning problems. Through a series of methods, it successfully reduces the dimensionality of image recognition problems with huge amounts of data, and finally makes them Can be trained. A typical convolutional neural network is composed of a convolutional layer, a pooling layer, and a fully-linked layer. The most important of these is the convolutional layer. In our model, the full link layer is discarded, and only the convolutional layer and the pooling layer are retained

Convolutional neural network is a deep neural network. Convolutional neural networks are more complex in structure, mainly composed of input layer, convolution layer, pooling layer, fully connected layer, output layer, etc. [7]. This paper mainly uses the convolutional layer and the pooling layer to extract features from the relevant data of Leiyun. In one-dimensional convolution, the role of convolution can be understood as extracting the translational features of the data in a certain direction. In the analysis of time series, it is expressed as extracting the features of the sequence. The essence of the convolution operation here is cyclic product and addition and, its mathematical expression is as follows:

$$y(k) = h(k) * u(k) = \sum_{i=0}^{N} h(k-i)u(i) \quad (3)$$

Among them $y, h, u$ are sequences, $k$ is the number of convolutions, and $N$ is the length of $u$ .

3.2 Long Short-Term Memory

Long Short-Term Memory unit is shown in the fig 1. The information transmission process of the hidden layer of the LSTM network is shown in the fig 2.

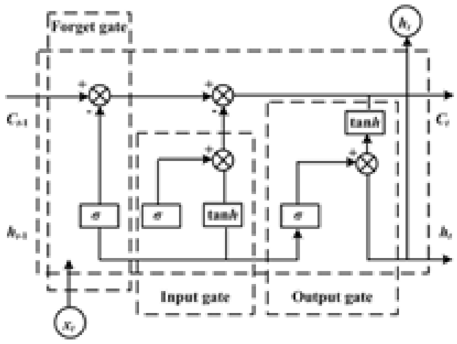


Fig. 1 LSTM memory cell structure diagram

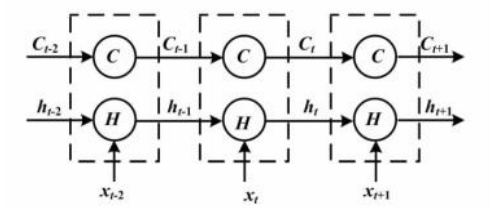

Fig. 2 LSTM hidden layer information transfer process diagram

Long Short-Term Memory network effectively mines the timing dependency of information in the time series by adding forget gates, input gates, and output gates in the hidden layer.

At each moment, the LSTM unit receives the current data input $x_j$ and the previous implicit state $h_{i-1}$ and the memory cell state $C_{t-1}$ through three gates. Then the calculation process of LSTM is as follows.[8]

The forget gate helps LSTM decide which information will be deleted from the memory cell state:

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{4}$$

LSTM uses input gate $i_t$ to determine the new information to be stored in the new unit state $C_t$ . The calculation process is as follows

$$\begin{aligned} i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \\ f_t &= \varnothing(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \\ C_t &= C_{t-1}f_t + g_t i_t \end{aligned} \tag{5}$$

In the formula: $g_t$ is the candidate value added to the new unit state; $C_{t-1}f_t$ is used to determine how much information will be forgotten from $C_{t-1}$ ; $g_t i_t$ determines how much information is added to the new unit state $C_t$ is the bitwise multiplication of elements in the vector.[9]

The process of calculating using output gate is as follows

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_0)$$
$$h_t = o_t \varnothing(C_t) \quad (6)$$

In the formula: $\sigma, \varnothing$ are the sigmoid activation function and tanh activation function respectively, $W_{fx}$, $W_{fh}$, $W_{ix}$, $W_{ih}$, $W_{gx}$, $W_{gh}$, $W_{ox}$, $W_{oh}$ is the weight matrix of the forget gate, input gate, input node, output gate, input $x_t$ and the previous hidden state $h_{t-1}$; $b_f$, $b_t$, $b_g$, $b_0$ is the corresponding bias; $f_t$, $i_t$, $g_t$, $o_t$, $C_t$, $h_t$ is the forget gate, input gate, input node, The output result of output gate, memory cell state and hidden state.

### 3.3 Input selection

The lightning data recorded by the LLS system includes the date and time of the lightning occurrence (including year, month, day, hour, minute), microseconds of the lightning strike, latitude, longitude, lightning current intensity and polarity, and the number of lightning return strikes [14]. By filtering and sorting these data, valid input values can be obtained.

Fig 3 and Fig 4 shows the number of mines in different months and time periods in 2018. Due to the difference in rainfall in different months, the number of thunder and lightning is obviously affected.

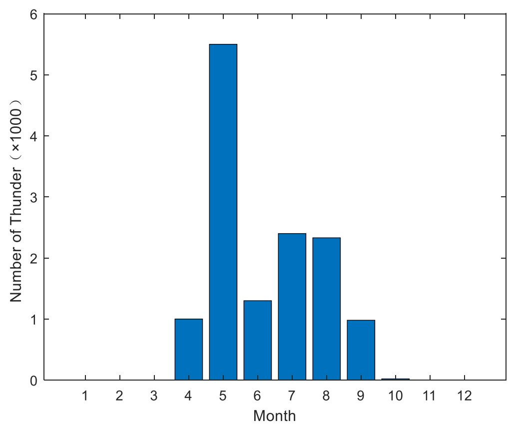


Fig. 3 Month and number of lightning

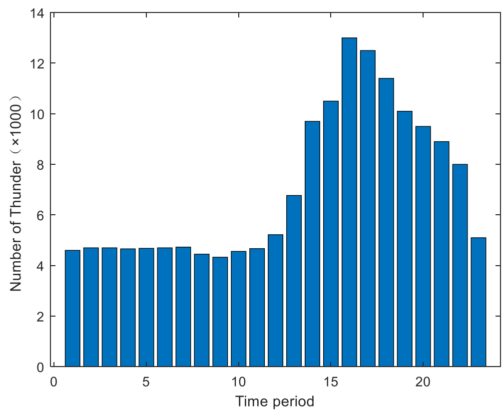


Fig.4Time and number of lightning

In summary, the algorithm flow of the thundercloud prediction model based on spatial density clustering and CNN-LSTM is shown in Figure 5.

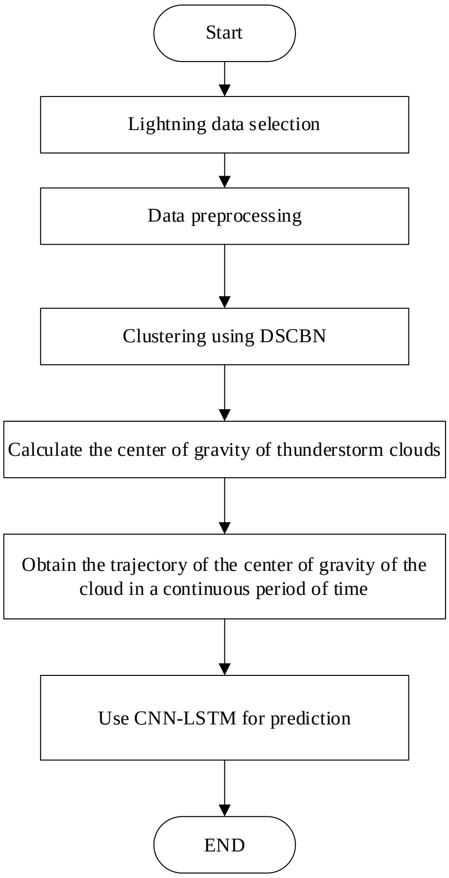


Fig. 5 Forecast algorithm flow

## 4. Case analysis

The author uses MATLAB2019b as the program development environment, and uses the data of the lightning positioning system of a certain regional power supply bureau from 2014 to 2018 for five years. Select August when the rainfall and the amount of lightning are abundant as an example for analysis, use the data from 2014 to 2017 as the training sample of the CNN-LSTM algorithm, and intercept a thunderstorm activity verification from 17:45-18:30 on July 25, 2018 Algorithm effect.

First, select the lightning strike data between 17:45 and 18:00 as the input data. According to the theory, divide them equally at 5 min intervals to obtain lightning strike data in 3 consecutive time periods. Then, the obtained data in each time period is clustered by spatial density, and the isolated and isolated lightning points are screened to form the thunderstorm cloud cluster at that time. The lightning falling in each interval before and after clustering is shown in Fig 6. Use equation (2) to find the center of gravity of the obtained thunderstorm cloud cluster, and the obtained values are shown in Table 1.

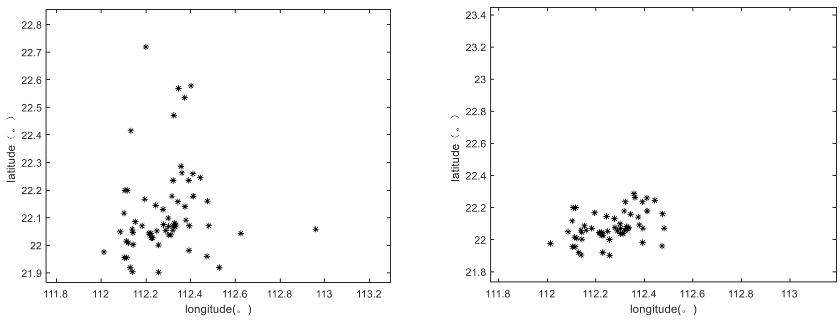


Fig. 6 Before and after clustering of lightning points

Table 1 Calculation result of cluster cloud center of gravity

| Forecast period | Center of gravity | |
|---|---|---|
| | longitude | latitude |
| 17:45-17:50 | 112.315 | 22.156 |
| 17:50-17:55 | 112.343 | 22.159 |
| 17:55-18:00 | 112.379 | 22.152 |

Then use the CNN-LSTM algorithm to predict the trajectory of the centroid movement. Since the dimensions of the input data are different, it is necessary to transform the different data to the same dimension to prevent the weight imbalance between the data. Using the maximum-minimum norm method, linear transformation is performed on the data without changing the law of data distribution.

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \left( x_{\max}^{new} - x_{\min}^{new} \right) + x_{\min}^{new} \tag{7}$$

In the formula, $x_{\max}$ , $x_{\min}$ are the maximum and minimum values of the original field respectively, which are mapped to the new interval $\left[ x_{\max}^{new} - x_{\min}^{new} \right]$ for normalization, generally the interval is [0, 1]

The date, time, and the latitude and longitude values of 3 locations are used as input variables, and CNN-LSTM is used for predictive analysis. And introduce SVM, LSTM and CNN-LSTM algorithm results to compare with the actual center of gravity.

Fig 7 shows the comparison between the three calculation methods and the actual lightning cloud cluster centers in the and after periods.

It can be seen from Figure 7 that compared with SVM and LSTM, the prediction accuracy of the CNN-LSTM algorithm is higher. Analyze the error between the predicted cloud cluster centroid calculated by each method and the actual centroid, and compare the accuracy of each algorithm under different time and average states. The results are shown in Table 2. It can be seen that the CNN-LSTM algorithm is correcting the thundercloud cluster. The average error in the center of gravity prediction is smaller, and the prediction accuracy is better than the other two algorithms.

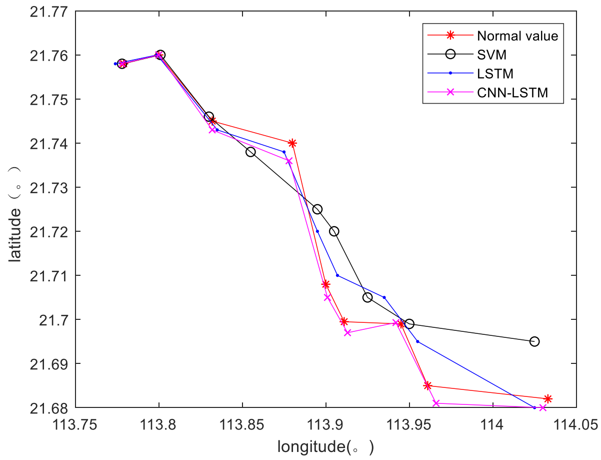


Fig .7 Comparison of different algorithms and actual value curves

Table2: Comparison of prediction errors of various prediction methods

| Forecast period | Absolut error /(°) | | | | | |
|---|---|---|---|---|---|---|
| | SVM | | LSTM | | CNN-LSTM | |
| | longitude | latitude | longitude | latitude | longitude | latitude |
| 18:00-18:05 | 0.014 | 0.002 | 0.009 | 0.003 | 0.006 | 0.003 |
| 18:05-18:10 | 0.006 | 0.023 | 0.013 | 0.012 | 0.009 | 0.002 |
| 18:10-18:15 | 0.033 | 0.018 | 0.006 | 0.008 | 0.003 | 0.005 |
| 18:15-18:20 | 0.026 | 0.006 | 0.015 | 0.004 | 0.011 | 0.006 |
| 18:20-18:25 | 0.012 | 0.023 | 0.008 | 0.003 | 0.002 | 0.002 |
| 18:25-18:30 | 0.011 | 0.014 | 0.007 | 0.009 | 0.004 | 0.001 |
| Average | 0.017 | 0.014 | 0.009 | 0.007 | 0.006 | 0.002 |

## 5. Conclusion

This paper uses the combination of DBSCAN clustering method and CNN-LSTM to predict. DBSCAN can solve the spatial problem in the lightning prediction process, and CNN-LSTM mainly solves the problem of data timing. Compared with other algorithms, the method in this paper has achieved better results.

But there are still shortcomings to be improved

1) The forecast time can still be advanced. This article uses 15min before the start of

thunderstorm as the initial input condition. If it can be combined with other information such as atmospheric weather, it should be possible to advance the forecast time, so that the management and dispatcher can make preparations earlier.

2) The accuracy of the prediction results can be further improved in combination with other machine learning models.

## Acknowledgements

This paper was financially supported by The Innovation Fund of Postgraduate, Sichuan University of Science & Engineering(Y201915), Enterprise Informationization and Internet of Things Measurement and Control Technology Sichuan Province University Key Laboratory Open Project(2018WZY03)

## References

[1]WU Chuan-qi, DING Li, HE Heng-xin, et al. Spatial andtemporal distribution characteristics of lightning activityof Hunan Power Grid [J]. High Voltage Engineering, 2010,36(4): 932-938

[2]Meyer V K, Höller H, Betz H D. Automated thunderstorm tracking: utilization of three-dimensional lightningand radar data[J]. Atmospheric Chemistry and Physics, 2013, 13(10): 5 137-5 150.

[3] LAN Hong-ping, SUN Xiang-ming, LIANG Bi-ling, et al.An automatic tracking and recognition algorithm forthunderstorm cloud-cluster[J]. Meteorological Monthly,2009, 35(7): 101-111, 131.

[4]GAO Wen-shen, ZHANG Bo-wen, ZHOU Rui-xu, et al.Nowcasting of the thunderstorm trend based on datacollected by lightning location system[J]. Power SystemTechnology, 2015, 39(2): 523-529.

[5]CHEN Jia-hong, FENG Wan-xing, WANG Hai-tao, et al.Statistical method of lightning parameters[J]. High VoltageEngineering, 2007, 33(10): 6-10

[6] Juntian G, Shanqiang G, Wanxing F. A lightning motionprediction technology based on spatial clusteringmethod[C]. 7th Asia-Pacific International Conference on.IEEE, Chengdu, China, 2011

[7] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature,2015, 521(7553):436

[8]SS K，AYUSH K，BABU Ｒ V. DeepFix: A fully convolutionalneural network for predicting human eye fixations ［J］. IEEETransactions on Image Processing，2017，26( 9) : 4446-4456.

[9] He K, Zhang X, Ren S, et al. Deep Residual Learning forImage Recognition[J]. 2015:770-778.