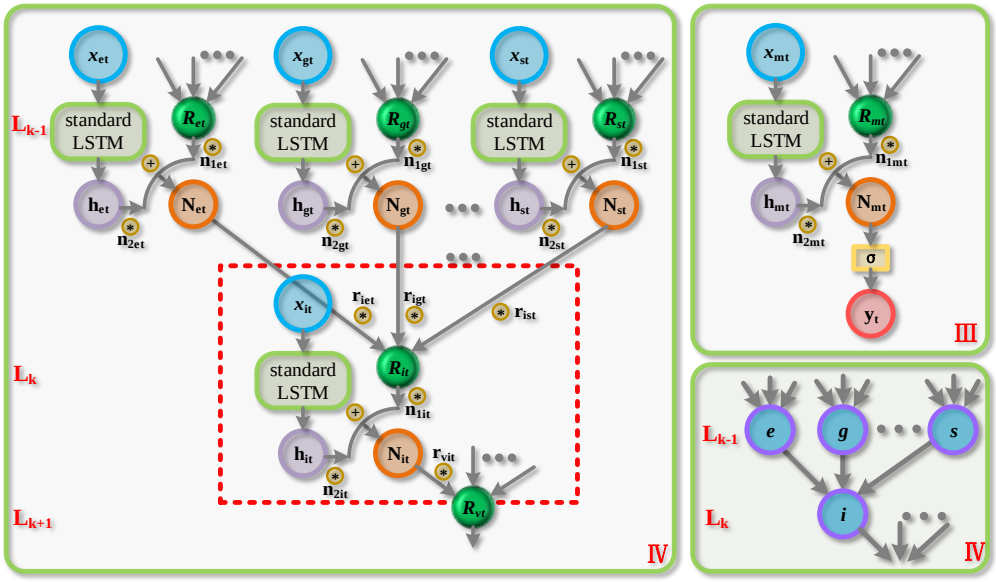

Fig. 1. Schematic diagram of NLSTM in the *t*-th period.

1、Forward propagation of NLSTM

Forward propagation of NLSTM in the *t*-th period are as follows:

(1) All nodes complete the forward propagation of the standard LSTM. Take node *i* as an example：

$$f_{it} = \sigma(\text{net}_{f,i,t}) = \sigma(w_{fh,i} \bullet h_{i,t-1} + w_{fx,i} \bullet x_{it} + b_{f,i}) \quad (1)$$

$$i_{it} = \sigma(\text{net}_{i,i,t}) = \sigma(w_{ih,i} \bullet h_{i,t-1} + w_{ix,i} \bullet x_{it} + b_{i,i}) \quad (2)$$

$$a_{it} = \tanh(\text{net}_{a,i,t}) = \tanh(w_{ah,i} \bullet h_{i,t-1} + w_{ax,i} \bullet x_{it} + b_{a,i}) \quad (3)$$

$$C_{it} = f_{it} * C_{i,t-1} + i_{it} * a_{it} \quad (4)$$

$$o_{it} = \sigma(\text{net}_{o,i,t}) = \sigma(w_{oh,i} \bullet h_{i,t-1} + w_{ox,i} \bullet x_{it} + b_{o,i}) \quad (5)$$

$$h_{it} = o_{it} * \tanh(C_{it}) \quad (6)$$

where $[w_{fh,i}, w_{fx,i}, b_{f,i}]$, $[w_{ih,i}, w_{ix,i}, b_{i,i}]$, $[w_{ah,i}, w_{ax,i}, b_{a,i}]$ and $[w_{oh,i}, w_{ox,i}, b_{o,i}]$ are weights and bias. f_{it}, i_{it}, o_{it} and a_{it} denote for the forget gates, the input gates, the output gates and the current information state. $C_{i,t-1}$ and $C_{i,t}$ represent for the cell state of the previous period and current period. $h_{i,t-1}$ and h_{it} stand for the standard LSTM outputs of the previous period and current period. $\text{net}_{\_,i,t}$ is intermediate variable. The symbol $\bullet$ indicates matrix multiplication and the symbol $*$ indicates multiplication between matrix elements. $\sigma(x)$ and $\tanh(x)$ are activation function of *Sigmoid* and *Tanh* [19].

(2) Forward propagation along the tree. Take node *i* as an example：

$$r_{ijt} = \sigma(\text{net}_{r,i,j,t}) = \sigma(w_{rh,i,j} \bullet h_{i,t-1} + w_{rx,i,j} \bullet x_{it} + b_{r,i,j}), (P_i \neq \varnothing) \tag{7}$$

$$R_{it} = \begin{cases} \sum_{j=1}^{len(P_i)} r_{ijt} * N_{P_{ij},t} & (P_i \neq \varnothing) \\ 0 & else \end{cases} \tag{8}$$

$$n_{1it} = \sigma(\text{net}_{n1,i,t}) = \sigma(w_{n1h,i} \bullet h_{i,t-1} + w_{n1x,i} \bullet x_{it} + b_{n1,i}) \tag{9}$$

$$n_{2it} = \sigma(\text{net}_{n2,i,t}) = \sigma(w_{n2h,i} \bullet h_{i,t-1} + w_{n2x,i} \bullet x_{it} + b_{n2,i}) \tag{10}$$

$$N_{it} = n_{1it} * R_{it} + n_{2it} * h_{it} \tag{11}$$

where $[w_{n1h,i}, w_{n1x,i}, b_{n1,i}]$, $[w_{n2h,i}, w_{n2x,i}, b_{n2,i}]$ and $[w_{rh,i,j}, w_{rx,i,j}, b_{r,i}]$ are weights and bias. P_i is the set of child nodes of node i. P_{ij} is the *j*-th child node number of node i. Function *len()* is used to calculate the length of the set. Other variables have the same meaning as mentioned before.

(3) Prediction:

$$y_t = \sigma(z_t) = \sigma(w_y \bullet N_{mt} + b_y) \tag{12}$$

where $[w_y, b_y]$ are weights and bias. z_t is intermediate variable.

2、Back propagation of NLSTM

In the NLSTM, $[w_{fh,i}, w_{fx,i}, b_{f,i}]$, $[w_{ih,i}, w_{ix,i}, b_{i,i}]$, $[w_{ah,i}, w_{ax,i}, b_{a,i}]$, $[w_{oh,i}, w_{ox,i}, b_{o,i}]$, $[w_{n1h,i}, w_{n1x,i}, b_{n1,i}]$, $[w_{n2h,i}, w_{n2x,i}, b_{n2,i}]$, $[w_{rh,i,j}, w_{rx,i,j}, b_{r,i}]$, and $[w_y, b_y]$ are weights and bias. The purpose of back propagation is to solve their gradient $[\delta w_{fh,i}, \delta w_{fx,i}, \delta b_{f,i}]$, $[\delta w_{ih,i}, \delta w_{ix,i}, \delta b_{i,i}]$, $[\delta w_{ah,i}, \delta w_{ax,i}, \delta b_{a,i}]$, $[\delta w_{oh,i}, \delta w_{ox,i}, \delta b_{o,i}]$, $[\delta w_{n1h,i}, \delta w_{n1x,i}, \delta b_{n1,i}]$, $[\delta w_{n2h,i}, \delta w_{n2x,i}, \delta b_{n2,i}]$, $[\delta w_{rh,i,j}, \delta w_{rx,i,j}, \delta b_{r,i}]$ and $[\delta w_y, \delta b_y]$. The backpropagation of the error in the NLSTM has two directions, one is to propagate from the back period to the previous period, and the other is to propagate from the root node to the leaf node. The process and formula of back propagation are derived as follows.

First, the squared error function is defined as the objective function to be optimized.

$$E_t = \frac{1}{2}(\mathrm{y}_t - \mathrm{Y}_t)^2 \tag{13}$$

where E_t denotes the error in the *t*-th period. y_t and Y_t are predictions and observations in the *t*-th period, respectively.

Then, the gradient of each variable in the output layer is calculated. The source of error for y_t is E_t, and the source of error for z_t is y_t. The source of error for $[w_y, b_y]$ is z_t.

$$\delta y_t = \frac{\partial E_t}{\partial \mathrm{y}_t} = \mathrm{y}_t - \mathrm{Y}_t \tag{14}$$

$$\delta z_t = \frac{\partial E_t}{\partial z_t} = \frac{\partial E_t}{\partial \mathrm{y}_t}\frac{\partial \mathrm{y}_t}{\partial z_t} = \delta y_t * [y_t * (1 - y_t)] \tag{15}$$

$$\delta w_y = \frac{\partial E_t}{\partial w_y} = \frac{\partial E_t}{\partial z_t}\frac{\partial z_t}{\partial w_y} = \delta z_t \cdot N_{mt} \tag{16}$$

$$\delta b_y = \frac{\partial E_t}{\partial b_y} = \frac{\partial E_t}{\partial z_t}\frac{\partial z_t}{\partial b_y} = \delta z_t \cdot 1 = \delta z_t \tag{17}$$

Next, the gradient of each variable along the tree is calculated. T represents the total number of periods. The source of error for N_{it} is $R_{Bi,t}$. B_i is the parent node number of node i. For the root node, the source of error for N_{it} is z_t. The source of error for R_{it} is N_{it}. The sources of error for h_{it} are N_{it}, $net_{n1,i,t+1}$, $net_{n2,i,t+1}$, $net_{o,i,t+1}$, $net_{a,i,t+1}$, $net_{i,i,t+1}$, $net_{f,i,t+1}$ and $net_{r,i,j,t+1}$. The source of error for o_{it} is h_{it}. The sources of error for C_{it} are h_{it} and $C_{i,t+1}$. The source of error for i_{it}, f_{it} and a_{it} are C_{it}. The source of error for n_{1it} and n_{2it} are N_{it}. The source of error for $net_{a,i,t}$ is a_{it}. The source of

error for $net_{G,i,t}$ is G_{it}. G can be $[i, f, o, n_1, n_2]$. The source of error for r_{ijt} is R_{it}. The source of error for $net_{r,i,j,t}$ is r_{ijt}. The sources of error for $w_{Lh,i}$, $w_{Lx,i}$ and $b_{L,i}$ are $net_{L,i,t}$.

$$\delta N_{it} = \frac{\partial E_t}{\partial N_{it}} = \begin{cases} \frac{\partial E_t}{\partial z_t}\frac{\partial z_t}{\partial N_{mt}} = \delta z_t \bullet w_y & i = m \\ \frac{\partial E_t}{\partial R_{Bi,t}}\frac{\partial R_{Bi,t}}{\partial N_{it}} = \delta R_{Bi,t} * r_{Bi,j,t} & i \neq m \end{cases} \tag{18}$$

$$\delta R_{it} = \frac{\partial E_t}{\partial R_{it}} = \frac{\partial E_t}{\partial N_{it}}\frac{\partial N_{it}}{\partial R_{it}} = \delta N_{it} * n_{1it} \tag{19}$$

$$\begin{cases} \delta h_{it} = \delta h_{it}^1 + \delta h_{it}^2 + \delta h_{it}^3 \\ \delta h_{it}^1 = \delta N_{it} * n_{2it} \\ \delta h_{it}^2 = \delta net_{n1,i,t+1} \bullet w_{n1h,i} + \delta net_{n2,i,t+1} \bullet w_{n2h,i} + \delta net_{o,i,t+1} \bullet w_{oh,i} \\ \qquad + \delta net_{a,i,t+1} \bullet w_{ah,i} + \delta net_{i,i,t+1} \bullet w_{ih,i} + \delta net_{f,i,t+1} \bullet w_{fh,i} \qquad t \neq T \\ \delta h_{it}^3 = \sum_{j=1}^{len(P_i)} \delta net_{r,i,j,t+1} \bullet w_{rh,i,j} \qquad t \neq T \quad and \quad P_i \neq \varnothing \end{cases} \tag{20}$$

$$\delta o_{it} = \frac{\partial E_t}{\partial o_{it}} = \frac{\partial E_t}{\partial h_{it}}\frac{\partial h_{it}}{\partial o_{it}} = \delta h_{it} * \tanh(C_{it}) \tag{21}$$

$$\delta C_{it} = \frac{\partial E_t}{\partial C_{it}} = \begin{cases} \frac{\partial E_t}{\partial h_{it}}\frac{\partial h_{it}}{\partial C_{it}} = \delta h_{it} * o_{it} * [1 - \tanh^2(C_{it})] & t = T \\ \frac{\partial E_t}{\partial h_{it}}\frac{\partial h_{it}}{\partial C_{it}} + \frac{\partial E_t}{\partial C_{i,t+1}}\frac{\partial C_{i,t+1}}{\partial C_{it}} = \delta h_{it} * o_{it} * [1 - \tanh^2(C_{it})] + \delta C_{i,t+1} * f_{i,t+1} & t \neq T \end{cases}$$

(22)

$$\delta i_{it} = \frac{\partial E_t}{\partial i_{it}} = \frac{\partial E_t}{\partial C_{it}}\frac{\partial C_{it}}{\partial i_{it}} = \delta C_{it} * a_{it} \tag{23}$$

$$\delta f_{it} = \frac{\partial E_t}{\partial f_{it}} = \frac{\partial E_t}{\partial C_{it}}\frac{\partial C_{it}}{\partial f_{it}} = \delta C_{it} * C_{i,t-1} \tag{24}$$

$$\delta a_{it} = \frac{\partial E_t}{\partial a_{it}} = \frac{\partial E_t}{\partial C_{it}}\frac{\partial C_{it}}{\partial a_{it}} = \delta C_{it} * i_{it} \tag{25}$$

$$\delta n_{1it} = \frac{\partial E_t}{\partial n_{1it}} = \frac{\partial E_t}{\partial N_{it}}\frac{\partial N_{it}}{\partial n_{1it}} = \delta N_{it} * R_{it} \tag{26}$$

$$\delta n_{2it} = \frac{\partial E_t}{\partial n_{2it}} = \frac{\partial E_t}{\partial N_{it}}\frac{\partial N_{it}}{\partial n_{2it}} = \delta N_{it} * h_{it} \tag{27}$$

$$\delta net_{a,i,t} = \frac{\partial E_t}{\partial net_{a,i,t}} = \frac{\partial E_t}{\partial a_{it}}\frac{\partial a_{it}}{\partial net_{a,i,t}} = \delta a_{it} * [1 - a_{it}^2] \tag{28}$$

$$\delta net_{G,i,t} = \frac{\partial E_t}{\partial net_{G,i,t}} = \frac{\partial E_t}{\partial G_{it}}\frac{\partial G_{it}}{\partial net_{G,i,t}} = \delta G_{it} * [G_{it} * (1 - G_{it})] \qquad G = [i, f, o, n_1, n_2] \tag{29}$$

$$\delta r_{ijt} = \frac{\partial E_t}{\partial r_{ijt}} = \frac{\partial E_t}{\partial R_{it}} \frac{\partial R_{it}}{\partial r_{ijt}} = \delta R_{it} * N_{P_{ij},t} \tag{30}$$

$$\delta net_{r,i,j,t} = \frac{\partial E_t}{\partial net_{r,i,j,t}} = \frac{\partial E_t}{\partial r_{ijt}} \frac{\partial r_{ijt}}{\partial net_{r,i,j,t}} = \delta r_{ijt} * [r_{ijt} * (1 - r_{ijt})] \tag{31}$$

$$\begin{cases} \delta w_{Lh,i} = \dfrac{\partial E_t}{\partial w_{Lh,i}} = \dfrac{\partial E_t}{\partial net_{L,i,t}} \dfrac{\partial net_{L,i,t}}{\partial w_{Lh,i}} = \delta net_{L,i,t} \bullet h_{i,t-1} \\ \delta w_{Lx,i} = \dfrac{\partial E_t}{\partial w_{Lx,i}} = \dfrac{\partial E_t}{\partial net_{L,i,t}} \dfrac{\partial net_{L,i,t}}{\partial w_{Lx,i}} = \delta net_{L,i,t} \bullet x_{it} \qquad L = [a, i, f, o, n_1, n_2, r] \\ \delta b_{L,i} = \dfrac{\partial E_t}{\partial b_{L,i}} = \dfrac{\partial E_t}{\partial net_{L,i,t}} \dfrac{\partial net_{L,i,t}}{\partial b_{L,i}} = \delta net_{L,i,t} \end{cases} \tag{32}$$