IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 23, 2017, accepted October 24, 2017, date of publication November 23, 2017,
date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2771798

Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network

HONGXIN XUE[1,2], YANPING BAI[2], HONGPING HU[2], AND HAIJIAN LIANG[3]

[1]School of Information and Communication Engineering, North University of China, Taiyuan 030051, China
[2]Department of Mathematics, School of Science, North University of China, Taiyuan 030051, China
[3]National Key Laboratory for Electronic Measurement Technology, Key Laboratory of Instrumentation Science and Dynamic Measurement Ministry of Educations, School of Information and Communication Engineering, North University of China, Taiyuan 030051, China

Corresponding author: Yanping Bai (baiyp666@163.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 61774137, in part by the Shanxi Natural Science Foundation under Grant 201701D22111439 and Grant 201701D221121, and in part by the Shanxi Scholarship Council of China under Grant 2016-088.

ABSTRACT In this paper, a series of models were established, and based on the Google Flu Trends (GFT) data and Centers for Disease Control (CDC) data. The models include the GFT regression model (model 1), the weighted GFT regression model (model 2), the GFT + CDC regression model (model 3), the CDC regression model (model 4), and the weighted CDC regression model (model 5). All models were utilized to predict and assess influenza activity across ten regions of the United States. The least squares and backpropagation neural network based on the genetic algorithm are used to fit the model parameters, and the error and historical sample fitting accuracy of each model are compared. The results show that models 4 and 5 are superior to other models. To optimize the prediction model, the seasonal characteristics of influenza incidence were investigated, and flu-prediction models for the high-flu season and low-flu season were established. The experimental results show that the prediction model of seasonal influenza is superior to the non-seasonal model. The influenza-like illness values predicted by the seasonal flu model are consistent with information provided by the CDC, suggesting that the results accurately reflect influenza epidemic characteristics and can thus be readily applied for the prevention and control of influenza.

INDEX TERMS CDC, GFT, GA-BP neural network, influenza prediction.

I. INTRODUCTION

Influenza (flu) was the first infectious respiratory disease to be monitored globally. It is a viral disease that crosses regions quickly via interpersonal communications. The disease is highly lethal among certain populations. It is difficult to make a laboratory diagnosis for each suspected influenza patient, so monitoring of influenza-like illness (ILI) often serves as a substitute for influenza surveillance [1]. ILI criteria comprise the case definition recommended by the World Health Organization (WHO) and the National Influenza Center (NIC); they include a body temperature greater than or equal to 38 °C accompanied by cough, pharyngeal pain, body pain, and other symptoms with a lack of another simultaneous laboratory diagnosis. According to the WHO, yearly flu epidemics cause 5-15% of the world's population to suffer from respiratory diseases, 3 million to 5 million cases of severe illness, and 250,000 to 500,000 deaths. The United States sustains economic losses of up to 71-167 billion USD due to the flu each year [2]. The prevention and control of influenza are critically dependent on timely monitoring of flu outbreaks. Early warnings can help to control the spread of the disease, significantly reducing the number of cases and the loss caused by an epidemic. To this effect, accurate flu monitoring has important practical significance to social stability, economic development, and the health and safety of vast swaths of the globe's population [3].

Researchers, both domestic and foreign, have investigated the surveillance and forecasting of influenza activity, typically by establishing prediction models for ILI epidemic trends. Li *et al.* [4], for example, built a predictive model to monitor the Chinese flu by using internet search data. They found that historical data can be utilized to predict influenza trends and that Internet search data can be used to improve the accuracy of current casting of predictions for the flu; the two types of data can be utilized in tandem to achieve favourable prediction effects. Researchers have

explored numerous methods for using network activities to implement flu surveillance, including patient counselling-centre hotlines [5], relevant site views [6], [7], Python-based Sina micro-blog crawler information [8], and Over The Counter (OTC) medication sales [9]. Research has also shown that official ILI data can be almost perfectly replaced by the flu-related model based on Twitter posts [10]–[15]. A variety of services, such as MappyHealth and Sickweather, can be used to gather data in real time, yielding a reliable assessment of the extent of damage caused by the flu.

The Google Flu Trends (GFT) system, launched in 2008, monitors and aggregates US search data [16] to estimate flu epidemics and compare its predictions against the Centers for Disease Control (CDC) data. A few weeks prior to the influenza A (H1N1) outbreak in 2009, Google engineers published a paper in *Nature* [17] to report that GFT successfully predicted the spread of H1N1, specific even to US regions and states, in a very timely manner. Google has indeed become a more efficient and timely indicator than official health organizations. In February 2013, *Nature* issued a report stating that the number of ILI clinic visits predicted by GFT was up to two times higher than that predicted by the CDC. The lead author David Lazer asserts that the disparity was caused by ''big data arrogance'' and algorithm variations [18].

Although numerous studies have shown that web search data are valuable, they are still far from adequate for replacing traditional methodologies for flu monitoring [19], [20]. Big data have substantial potential in regard to improving public health, but the numbers are misleading if background information is insufficient. Modern technologies cannot completely replace traditional epidemiological monitoring networks; they can only play a complementary role. In this study, we used a combination of CDC and GFT data to construct a novel combination forecasting model. Davidson *et al.* (2003) [19] suggested that flu levels in one region are strongly correlated with those in other regions, so the flu levels across several regions were integrated into a real-time CDC prediction model to construct an empirical network model.

American scholars, including Lowen *et al.* [21], have utilized countless guinea pigs under a variety of temperature conditions to study influenza activity. At 5°C, the propagation velocity of the flu virus significantly accelerates. At 20°C, the spread of the flu virus is relatively slow. At 30°C, the transmission of the virus nearly stops. The National Institutes of Health (NIH) researchers used a technique called magic angle spinning nuclear magnetic resonance to create detailed images of how the virus's outer membrane responded to variations in temperature. The virus's outer membrane was composed chiefly of molecules known as lipids, explained by Polozov *et al.* [22] and Ezzine [23]. They found that at winter temperatures, the virus's outer covering, or envelope, hardens to a rubbery gel that shields the virus as it passes from person to person. With warmer weather, however, the protective gel melts to a liquid phase. But this liquid phase, apparently, isn't tough enough to protect the virus against the elements, and so the virus loses its ability to spread from person to person. These observations are well supported by the fact that influenza epidemics more commonly occur in winter and spring months than in the rest of the year. In this study, we analysed the number of ILIs across ten regions of the United States and divided the year into a high-flu season and a low-flu season to construct the improved flu prediction model.

The introduction of neural networks into the prediction of influenza is a hot topic in biomedical information processing. The neural network, an intelligent recognition algorithm, is commonly used in various fields. It is fault tolerant and provides association, inference, memory, self-adaptation, self-learning, and the ability to manage complex multi-modal data [24]. The back-propagation (BP) neural network has played an important role in the development of artificial neural networks and is currently one of the most widely used networks in this field [25]. In the present study, we applied a BP neural network based on the genetic algorithm (GA-BP) to fit the parameters of the flu-prediction regression model and compared the fitting effect with that of Least Squares (LS), as discussed below.

In this paper, the following research work has been carried out. First, multiple regression models, namely, the GFT regression model (model1), weighted GFT regression model (model2), GFT+CDC regression model (model3), CDC regression model (model4) and weighted CDC regression model (model5), are established. Second, the LS and GA-BP methods are used to fit the model parameters, and the results are compared. Finally, according to the seasonal characteristics of influenza incidence, we divide the year into a high-flu season and a low-flu season to establish the quarterly seasonal flu prediction model.

II. METHODS AND MODELS

A. GFT REGRESSION MODEL

The GFT model, tracks ILIs in whole populations based on Google search query data [26]. Its working principle is that if a person is suffering from the flu, he or she is likely to search the Internet for information related to the disease. By extracting search-engine keywords associated with the flu, such as ''cold'', ''sore throat'', ''fever'' and other symptoms of disease, and analysing the data, regional influenza epidemic situations can be estimated fairly accurately [27]. Scores are assigned to the database terms commonly searched by 50 million users by fitting them to the model. The N most commonly searched terms are selected. Then, the ability of the model to predict ILI can be evaluated based on the numbers of times these terms are searched. When $N = 45$, for example, the model yields prediction results that are in close agreement with the CDC ILI data [28], [29]. We used the GFT data published by the Google Flu Monitoring System to construct the regression models of Eq.(1). Davidson *et al.* [19] constructed an empirical network model using GFT data, refer to (2). Below, we will compare the predictive effects of

these two models.

$$ILI_{i,t} = \sum_{k=1}^{P} \chi_k X_{i,t-k} + \tau_t. \quad (1)$$

$$ILI_{i,t} = \beta_1 X_{i,t} + \sum_{j\neq i, j=1}^{N} \lambda_j \omega_{i,j} X_{j,t} + \nu_t. \quad (2)$$

where $ILI_{i,t}$ is the number of CDC ILIs in the i-th region in week t; $X_{i,t}$ is the number of GFT ILIs in the i-th region in week t; $\omega_{i,j}$ is the weight of the relationship between region i and region j, for which authors use the cross-correlation between the laboratory-confirmed influenza cases for each pair of regions (i, j) in the previous year (the greater the correlation, the greater the value of $\omega_{i,j}$); P is the lagged order of the dependent variable $X_{i,t-k}$, (the experimental results show that prediction effect of the model is the best when $P = 4$); N is the number of regions ($N = 10$); χ_k, λ_j are the coefficient of models; and τ_t, ν_t are the residual of the model for week t.

B. CDC REGRESSION MODEL

The US CDC routinely counts weekly influenza-related cases among all outpatients in ten surveillance regions [30] and releases weekly influenza-transmission data from national and regional disease-surveillance systems. The sizable error in the GFT data and CDC data means that the re-use of GFT data to build an empirical network model will lead to secondary error and make the error is increasing. Therefore, we used the CDC data to establish the flu-prediction model. Because the spread and infection of flu are continuous in time, the data processed by the predictive model are also continuous in time. Therefore, when building the model, we consider the effect of the previous flu on the next moment. The predicted flu data are closest to the real CDC data when the lag time is 4 weeks($P = 4$); accordingly, we established the following regression model with a lag time of 4 weeks:

$$ILI_{i,t} = \sum_{k=1}^{P} \alpha_k ILI_{i,t-k} + \varepsilon_t. \quad (3)$$

where $ILI_{i,t-k}$ is the number of CDC ILIs in the i-th region in week $t - k$; α_k is the model coefficient; P is the same as those in Eq.(1)($P = 4$); and ε_t is the residual of the model.

The flu is mostly spread by contact among individuals. Thus, the relationship between any two regions of flu-affected populations is affected by the relationship between the people inhabiting these regions. If there is an outbreak of influenza in a region with frequent population movements, regions in close association are likely to be implicated. We integrated influenza-level information from across various regions into the real-time prediction model to establish the weighted CDC regression model. We used historical CDC data to obtain the correlation coefficient between each region as the weight of the relationship: the greater the correlation, the greater the $\omega_{i,j}$ value. Each region has nine relationship weights, $\omega_{i,j} \in [0, 1]$. The model formula is as follows:

$$ILI_{i,t} = \sum_{k=1}^{P} \beta_k ILI_{i,t-k} + \sum_{j\neq i, j=1}^{N} \lambda_j \omega_{i,j} ILI_{j,t} + \theta_t. \quad (4)$$

where the parameters $ILI_{i,t-k}, \beta_k, \lambda_j, P$, and θ_t are the same as those in Eq. (3); $\omega_{i,j}$ is the weight of the relationship between region i and region j; and N represents the number of regions included in the model; here, $N = 10$.

C. GFT+CDC REGRESSION MODEL

The CDC regression in the flu-prediction model uses historical ILI data to predict future trends. In the absence of a new mutation impacting the case, this method yields good predictions. Historical flu-outbreak information is indeed useful in predicting future flu activity, but the GFT regression in the model is still valuable in that it is better able to reflect disease outbreaks in real time. In short, combining the two models yields the more accurate possible predictions. We built the following combination forecasting model accordingly:

$$ILI_{i,t} = \sum_{k=1}^{P} \mu_k ILI_{i,t-k} + \sum_{m=1}^{P} \delta_m X_{i,t-m} + \sigma_t. \quad (5)$$

where the parameters $ILI_{i,t-k}, \mu_k, \delta_m, \sigma_t$ and P are the same as those in Eq. (3) and $X_{i,t-m}$ is the number of GFT ILIs in the i-th region in week $t - m$.

D. SEASONAL PREDICTION MODEL

Influenza is generally considered to have seasonal epidemic characteristics, with a higher rate of prevalence in winter and spring months compared to the rest of the year. Increased contact among individuals due to being confined indoors under inclement weather, as well as the body's weakened resistance to disease during environmental temperature fluctuations, are some of the reasons for variations in flu prevalence by season. As discussed above, we analysed the number of ILIs in ten regions of the United States. Fig.1 shows the ILI distribution across all regions from week 40 in 2003 to week 33 in 2015, for a total of 620 weeks, which shows that the flu indeed has a significant seasonal incidence peak. Each region's flu prevalence begins to rise at 40 weeks, peaks at weeks 48-52, then declines during weeks 1-5 of the following year and returns to normal after 16 weeks. We divided the year into a high-flu season (weeks 40-15) and a low-flu season (weeks 16-39) to establish the improved model.

E. GA-BP MODEL

The goal of LS regression analysis is to find a concrete form of functional expression and to establish and analyse the functional relationship between a response (dependent variable Y) and an important factor (independent variable X), that is, to obtain a suitable function of X with which to express Y. The purpose of neural networking based on regression analysis is to find a neural network model [31], not a specific mapping-function expression. It is through the network that the training

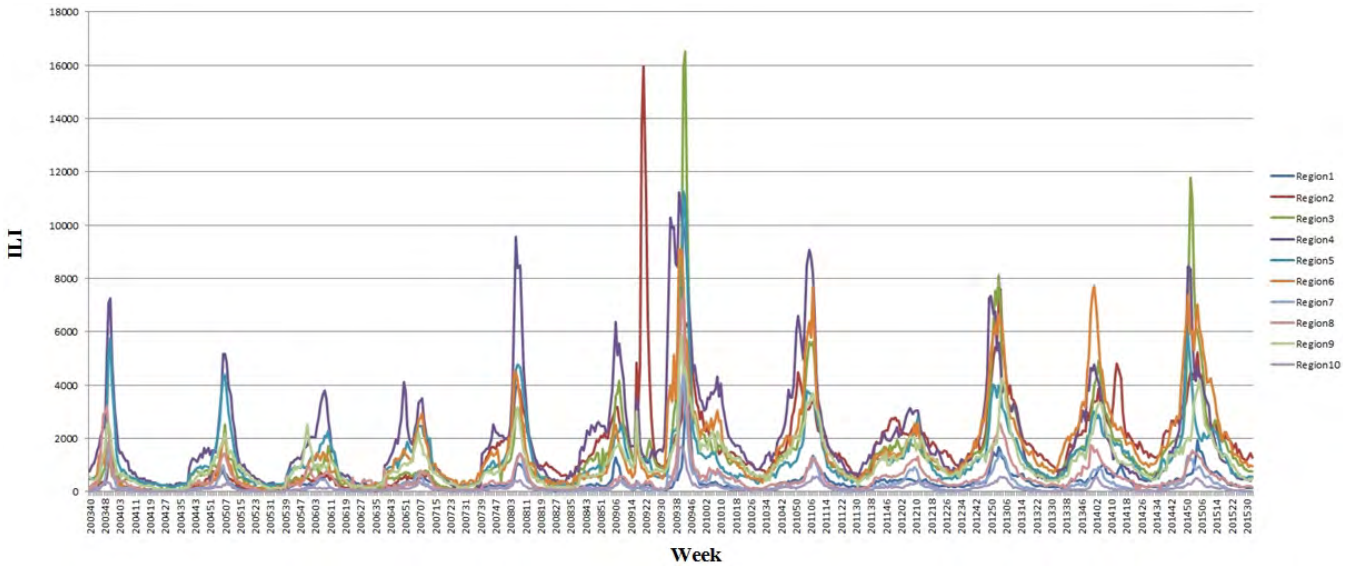


FIGURE 1. Distribution of ILIs from week 40 in 2003 to week 33 in 2015 in ten regions of the United States.

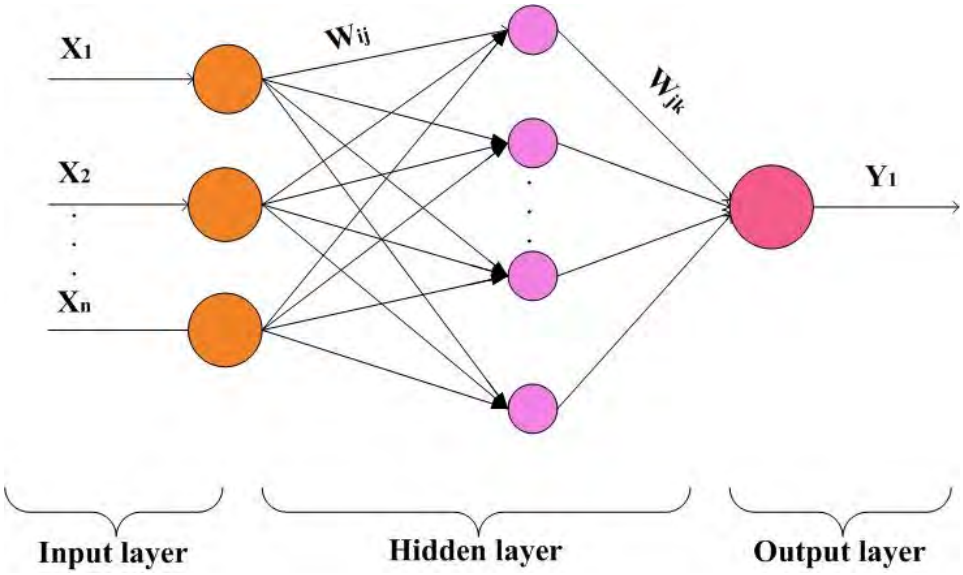


FIGURE 2. Topological structure of a BP Neural Network.

sample is learned; when the network training is complete, the network structure represents mapping $X \rightarrow Y$.

1) BP NEURAL NETWORK

A BP neural network is a multilayer feedforward network trained by an error inverse-propagation algorithm. The basic idea is to study the sample pair (input and expected output) and send the input of the sample to the BP neural network to calculate the actual output. If the error between the actual output and the expected output does not meet the accuracy requirement, the error is propagated backward to adjust the weight ω and the threshold b so that the error between the output of the network and the expected output is gradually reduced until the accuracy requirement is satisfied. The topology of the BP network is shown in Fig. 2.

In Fig. 2, $X_1, X_2, ..., X_n$ are the input values of the BP neural network, Y_1 is the predicted value of the BP neural network, and ω_{ij} and ω_{jk} are BP neural-network weights.

2) GA-BP MODEL

The BP neural network is one of the most widely applied artificial neural networks. Its structure is rigorous, it has strong operability, and it can realize highly non-linear mapping. It is disadvantaged by slow learning convergence, elusive network topology, and a tendency to fall into the local minimum [32]. The genetic algorithm can be applied to optimize the BP neural-network model to check several feasible solutions at the same time, which can speed up the convergence of the network and improve the prediction accuracy of the

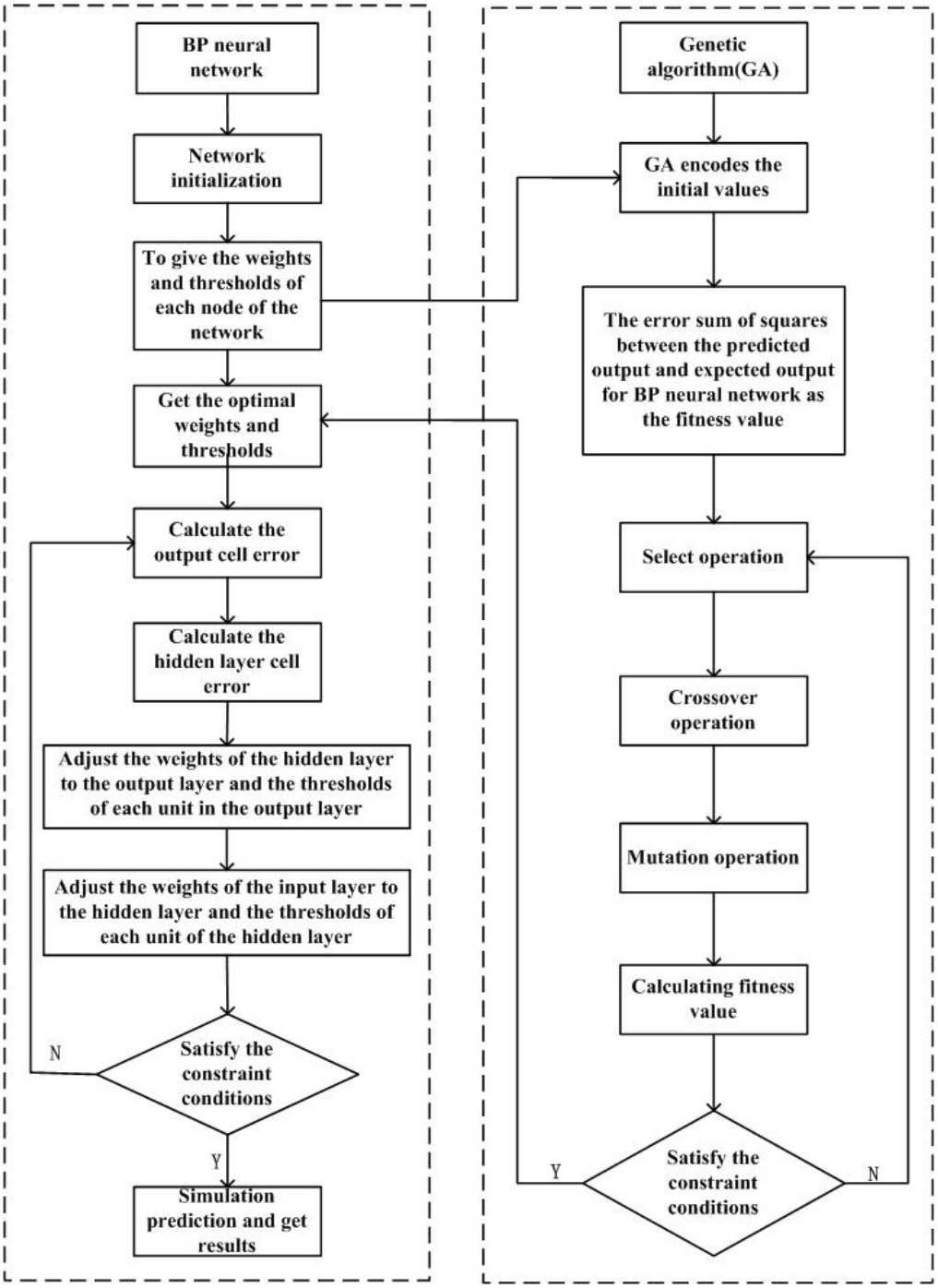


FIGURE 3. GA-BP neural network algorithm.

network model [33]. We fit the parameters of the proposed flu-prediction model based on GA-BP.

The specific steps of the genetic algorithm to optimize the BP neural network are as follows:

Step 1: Structure determination of the BP neural network. The individual coding length of the genetic algorithm is determined by the number of input and output parameters, which determine the structure of the BP neural network. If n is the number of nodes in the input layer, l is the number of nodes in the hidden layer, and m is the number of nodes in the output layer, the coding length is: $S = n \times l + l \times m + l + m$.

Step 2: Genetic-algorithm optimization. The genetic algorithm is used to optimize the weights and thresholds of the BP neural network. The individual fitness value is calculated by a fitness function. Then the genetic algorithm finds the optimal fitness value by selection, crossover, and mutation operations.

The genetic-algorithm-optimized BP neural-network elements include population initialization, the fitness function,

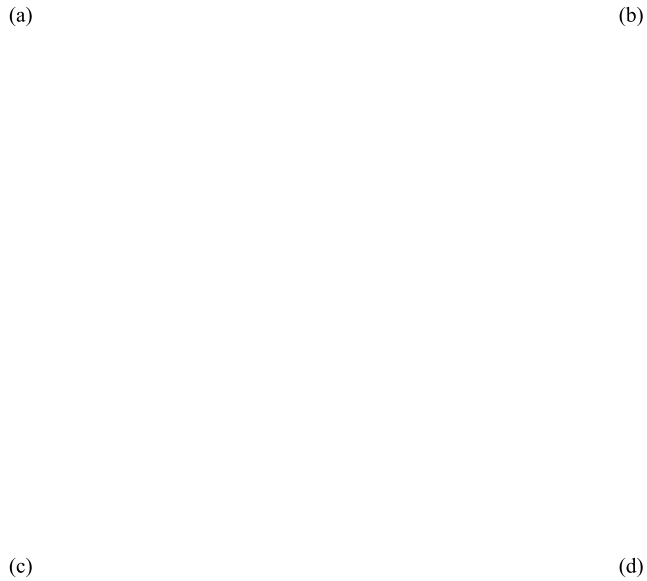


FIGURE 4. Prediction-error charts of region 3 and region 10. (a) LS test errors of Region 3. (b) GA-BP test errors of Region 3. (c) LS test errors of Region 10. (d) GA-BP test errors of Region 10.

selection operation, crossover operation, and mutation operation [34].

Step 2.1: Population initialization. The individual encoding method is real coding, in which each individual is a real string consisting of an input-layer and hidden-layer weight value, a hidden-layer threshold, a hidden-layer and output-layer weight value, and an output-layer threshold. The individual contains all the weights and thresholds of the neural network.

Step 2.2: Fitness function. The initial weights and thresholds of the BP neural network are obtained according to the individual. The trained BP neural network is used for prediction. The sum of the squared errors between the predicted output and expected output is taken as the individual fitness value F:

$$F = r[\sum_{i=1}^{m}(y_i - \overline{y_i})^2]. \tag{6}$$

where m is number of network output nodes; y_i is the expected output of the i-th node; $\overline{y_i}$ is the predicted output of the i-th node; and r is the coefficient.

Step 2.3: Selection operation. The selection operation of the genetic algorithm includes the random-traversal method, roulette method, tournament method, and other methods. The roulette method was chosen here, in which the individual is selected and transmitted to the next-generation population with some probability and the individual's probability is proportional to the size of the fitness. The selection probability p_i of each individual i is:

$$f_i = \frac{r}{F_i}. \tag{7}$$

$$p_i = \frac{f_i}{\sum_{j=1}^{M} f_j}. \tag{8}$$

where F_i is the fitness value of the i-th individual (a smaller fitness value is preferable). The reciprocal is calculated to determine the fitness value prior to individual selection. r is the coefficient, and M is the number of individuals.

Step 2.4: Crossover operation. Because the individual is used by the real-number code, the crossover-operation method uses the real-number crossover method. The intersection of the s-th chromosome g_s and the t-th chromosome g_t at position i is performed as follows:

$$\begin{cases} g_{si} = g_{si}(1-c) + g_{ti}c \\ g_{ti} = g_{ti}(1-c) + g_{si}c. \end{cases} \tag{9}$$

where c is a random number in [0, 1].

Step 2.5: Mutation operation. The j-th gene g_{ij} of the i-th individual is selected for mutation:

$$g_{ij} = \begin{cases} g_{ij} + (g_{ij} - g_{max}) * f(t) & u > 0.5 \\ g_{ij} + (g_{min} - g_{ij}) * f(t) & u \leq 0.5. \end{cases} \tag{10}$$

where g_{max} is the upper bound of gene g_{ij}; g_{min} is the lower bound of gene g_{ij}; $f(t) = u(1 - \frac{t}{T_{max}})^2$; u is a random number; t is the current iteration number; T_{max} is the maximum number of iterations; and u is a random number in [0, 1].

Step 3: BP neural-network prediction. The GA is used to obtain the optimal individual, which is used to determine the initial weights and thresholds of the BP network. The trained BP neural network then yields prediction results [35].

A flow chart of the GA-BP algorithm is shown in Fig. 3.

III. RESULTS

A. DATA COLLECTION

As discussed above, influenza activity data for the purposes of this study were gathered from both the CDC and the GFT [36], [37], and the experiment was conducted in MATLAB (R2014a). We selected GFT data and CDC data for 604 weeks from week 40 in 2003 to week 17 in 2015 to train each model, with week 18 to week 33 in 2015 as the test samples. The train data were utilized to fit the models, and the trained model was used to predict the ILI data in the test set. The predicted data were compared with the real ILI data to test the predictive abilities of the various models.

B. DATA PREPROCESSING

The data were normalized to prevent excessive disparity between the data and unwanted differences in dimension that would have affected prediction accuracy and to ensure that smaller output values were not obscured. We used MATLAB's mapminmax function to normalize the original input data, evenly distributing them in the range of [0, 1] via the following function:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}. \tag{11}$$

where x_i' is the parameter value of each sample input attribute after the normalization process; x_i is the parameter value of each sample input attribute before the normalization process; x_{max} is the maximum value in the original data sequence; and x_{min} is the minimum value in the original data sequence.

TABLE 1. Comparison of flu-prediction results.

Region 3	Method	MSE	$RMSE$	$MAPE(\%)$
Model 1	LS	8.5045e+05	0.5031	60.5858
	NET	6.2772e+05	0.3299	51.5190
Model 2	LS	4.9017e+05	0.3816	53.0634
	NET	1.5994e+05	0.0871	26.2857
Model 3	LS	3.8498e+04	0.0319	14.6974
	NET	7.0350e+04	0.0460	17.6407
Model 4	LS	1.7283e+04	0.0116	8.7504
	NET	1.3367e+04	0.0083	7.3531
Model 5	LS	1.2219e+04	0.0063	5.8859
	NET	2.0746e+04	0.0092	7.5220

Region 10	Method	MSE	$RMSE$	$MAPE(\%)$
Model 1	LS	1.6768e+03	2.1659	109.6323
	NET	1.4334e+03	0.7155	72.7346
Model 2	LS	2.1846e+03	2.9594	120.5290
	NET	1.8382e+03	1.3674	87.0039
Model 3	LS	2.8643e+02	0.3284	43.5995
	NET	8.3235e+02	0.4220	56.9223
Model 4	LS	5.4713e+02	0.3914	45.9301
	NET	5.0201e+02	0.1656	38.8700
Model 5	LS	2.7945e+02	0.1774	32.9315
	NET	5.4529e+02	0.2433	40.4312

To further compare and analyse the results of the various models, we used the mean square error (MSE), mean absolute percentage error ($MAPE$), and relative mean square error ($RMSE$):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_i})^2. \tag{12}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \mid \frac{y_i - \overline{y_i}}{\overline{y_i}} \mid \times 100\%. \tag{13}$$

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \overline{y_i}}{\overline{y_i}})^2. \tag{14}$$

where n is the number of samples; y_i is the number of actual ILIs; and $\overline{y_i}$ is the number of predicted ILIs.

C. MODEL ESTIMATION RESULTS

As discussed above, the training samples were used to fit the various model parameters, and then, each trained model was utilized to predict the numbers of ILIs in the ten regions of the United States from week 18 to week 33 in 2015. The MSE, $MAPE(\%)$ and $RMSE$ of the output were used for comparative analysis. Table 1 shows the prediction results for all five models in regions 3 and 10 as an examples. Models 1 through 5 correspond to the five models proposed in the second part of this paper, reference to formulas (1), (2), (5), (3) and (4). LS represent the prediction result of the

TABLE 2. Prediction results of model 6 and model 4 in the low-flu season.

Region	Method	Model	MSE	$RMSE$	$MAPE(\%)$
1	LS	Model 6	1.3698e+03	0.0278	14.4054
		Model 4	1.7134e+03	0.0345	15.2329
	NET	Model 6	1.1714e+03	0.0196	11.0896
		Model 4	1.1714e+03	0.0208	11.0896
2	LS	Model 6	6.1097e+04	0.0210	12.0705
		Model 4	6.2109e+04	0.0242	12.9026
	NET	Model 6	4.7780e+04	0.0157	10.9920
		Model 4	3.4461e+04	0.0141	10.2518
3	LS	Model 6	1.7283e+04	0.0100	8.0717
		Model 4	1.7283e+04	0.0116	8.7504
	NET	Model 6	1.1672e+04	0.0083	7.3531
		Model 4	1.3367e+04	0.0083	7.3531
4	LS	Model 6	7.8888e+03	0.0224	12.8327
		Model 4	1.4911e+04	0.0448	18.8753
	NET	Model 6	6.8554e+03	0.0144	9.5004
		Model 4	7.7445e+03	0.0144	9.[illegible]

TABLE 3. Prediction results of model 7 and model 5 in the low-flu season.

[illegible]

LS method, and NET is the prediction result of the GA-BP regression method.

The prediction results of model 1 and model 2 were analysed and compared. It was found that the weighted GFT regression model, which integrates the flu levels of the other nine regions into the real-time regression model of GFT, is superior to the model 1 in terms of prediction results. In other words, the relationships between regions do impact the influenza activity across the different regions.

The prediction results of model 2 and model 4 were analysed and compared. It was found that the CDC data yield a better prediction model than the GFT data, regardless of whether the method is LS or GA-BP, in regard to influenza prediction model *MSE* indicators. This is explained by the fact that the GFT prediction results in 2008 were highly correlated with the CDC ILI monitoring results, with a correlation coefficient of 0.97. As mentioned above, however, *Nature* later reported that GFT overestimated the incidence of influenza [18] in both the 2012 and 2013 flu seasons. In short, the GFT data diverge from the real (CDC) data, resulting in secondary error in model 2.

The prediction results of model 4 and model 3 are analysed and compared. Research shows that by adding GFT data to the historical ILI data, the model's forecasting effect has been improved. This indicates that the Google search data contain new information not contained in the historical data that reflects the variation of ILIs in the current period, and this new variation did not exist in the historical ILI data.

FIGURE 5. Prediction-error chart of Region 10 in the low-flu season. (a) LS test error of models 4 and 6. (b) GA-BP test error of models 4 and 6. (c) LS test error of models 5 and 7. (d) GA-BP test error of models 5 and 7.

The prediction results of model 4 and model 5 are analysed and compared. Research shows that the effect of the model is improved by adding the ILI data of other regions to the historical information. The reason is that influenza transmission depends mainly on the contact between people. In the ILI prediction, we can consider not only the role of historical sample data in this region but also the effects of region interactions on the spread of the disease.

The prediction results of model 3 and model 5 were also analysed and compared. It was found that model 5 outperforms model 3 whether parameterized using the LS method or GA-BP method.

From the above analysis, it can be concluded that among in the test output of the five models, the three evaluation indexes of model 5 are the best. The three evaluation indexes of the test outputs of GA-BP were better than the corresponding LS results, indicating that the neural-network method is superior to the LS method in predicting flu activity.

Prediction-error result charts for all five models in regions 3 and 10 are shown in Fig. 4 as an examples. It can be seen from Fig. 4 that models 4 and 5 are superior to other models for regions 3 and 10.

Altogether, the results indicated that the numbers of ILIs predicted by model 4 and model 5 are consistent with the numbers provided by the CDC. We next built improved prediction models based on models 4 and 5 for high-flu and low-flu seasons according to the epidemiological characteristics of the ten US regions we examined. A total of 354 samples were included in the low-flu season (weeks 16-45), and the first 338 samples were used as train set to training the models

TABLE 4. Prediction results of model 8 and model 4 in the high-flu season.

Region	Method	Model	MSE	$RMSE$	$MAPE(\%)$
1	LS	Model 8	5.2869e+04	0.0279	12.6102
		Model 4	5.8753e+04	0.0285	13.9717
	NET	Model 8	5.2437e+04	0.0249	11.6299
		Model 4	6.3147e+04	0.0329	14.2463
2	LS	Model 8	1.9292e+05	0.0140	9.0437
		Model 4	2.0448e+05	0.0151	10.6331
	NET	Model 8	1.8314e+05	0.0133	8.9936
		Model 4	1.8431e+05	0.0140	9.0009
3	LS	Model 8	2.8778e+06	0.0371	13.0164
		Model 4	2.0248e+06	0.0251	12.2260
	NET	Model 8	2.1951e+06	0.0284	12.9742
		Model 4	2.3122e+06	0.0314	14.0357
4	LS	Model 8	4.4630e+05	0.0121	9.9736
		Model 4	4.8902e+05	0.0287	10.4572
	NET	Model 8	2.3456e+05	0.0117	7.6174
		Model 4	2.3665e+05	0.0130	7.8298
5	LS	Model 8	1.0458e+05	0.0082	6.3222
		Model 4	1.2266E+05	0.0095	7.1931
	NET	Model 8	2.6648e+04	0.0044	5.3334
		Model 4	2.7370e+04	0.0050	5.6804
6	LS	Model 8	4.6055e+05	0.0136	8.9317
		Model 4	4.7737e+05	0.0139	9.1110
	NET	Model 8	3.5295e+05	0.0101	7.7937
		Model 4	4.4201e+05	0.0133	8.6867
7	LS	Model 8	1.4573e+04	0.0278	15.0514
		Model 4	1.5873e+04	0.0280	15.1022
	NET	Model 8	1.6476e+04	0.0342	16.1066
		Model 4	1.9368e+04	0.0392	17.0073
8	LS	Model 8	8.4598e+03	0.0082	7.3695
		Model 4	8.6145e+03	0.0080	7.7756
	NET	M0del 8	7.0181e+03	0.0075	7.0179
		Model 4	5.9494e+03	0.0063	6.7091
9	LS	Model 8	1.2442e+05	0.0145	10.0828
		Model 4	1.2608e+05	0.0147	10.1097
	NET	Model 8	1.7545e+05	0.0169	10.5709
		Model 4	2.3318e+05	0.0222	12.0356
10	LS	M0del 8	2.2003e+03	0.0388	16.1486
		Model 4	1.6844e+03	0.0368	15.1435
	NET	Model 8	2.0408e+03	0.0371	15.3886
		Model 4	2.2599e+03	0.0389	16.0747

TABLE 5. Prediction results of model 9 and model 5 in the high-flu season.

Region	Method	Model	MSE	$RMSE$	$MAPE(\%)$
1	LS	Model 9	1.2284e+05	0.1069	20.2422
		Model 5	8.1064e+05	0.1691	26.1526
	NET	Model 9	1.0201e+05	0.0533	18.1415
		Model 5	1.1679e+05	0.0567	19.6528
2	LS	Model 9	1.9139e+05	0.0162	9.3758
		Model 5	3.4359e+05	0.0234	13.5320
	NET	Model 9	4.2848e+05	0.0468	17.4915
		Model 5	6.7208e+05	0.0613	20.2906
3	LS	Model 9	2.1994e+06	0.0209	10.4448
		Model 5	2.5890e+06	0.0252	11.2263
	NET	Model 9	6.9547e+06	0.0816	22.5734
		Model 5	9.0993e+06	0.0984	23.4778
4	LS	Model 9	9.3325e+05	0.0862	23.0872
		Model 5	4.3059e+05	0.0325	13.9729
	NET	Model 9	9.5862e+05	0.1064	23.8393
		Model 5	1.0777e+06	0.1494	27.1560
5	LS	Model 9	1.0191e+05	0.0247	10.4297
		Model 5	1.2446e+05	0.0325	12.4580
	NET	Model 9	2.6801e+05	0.0408	18.2981
		Model 5	4.8035e+05	0.1494	20.4861
6	LS	Model 9	7.2260e+05	0.0211	9.7675
		Model 5	5.0581e+05	0.0147	9.1009
	NET	Model 9	1.2274e+06	0.0362	15.5452
		Model 5	1.3064e+06	0.0382	16.5560
7	LS	Model 9	7.2720e+03	0.0144	9.9614
		Model 5	7.3609e+03	0.0154	10.1123
	NET	Model 9	1.5447e+04	0.0388	15.5043
		Model 5	2.0526e+04	0.0736	20.9703
8	LS	Model 9	2.6449e+04	0.0236	11.1494
		Model 5	3.7521e+04	0.0342	16.4158
	NET	Model 9	3.4150e+04	0.0524	17.3828
		Model 5	4.3713e+04	0.0776	20.1685
9	LS	Model 9	7.1573e+04	0.0084	6.9812
		Model 5	1.1526e+05	0.0140	9.1859
	NET	Model 9	2.8979e+05	0.0248	12.5575
		Model 5	5.3023e+05	0.0475	17.5106
10	LS	M0del 9	3.0866e+03	0.0580	18.8201
		Model 5	2.0153e+03	0.0323	14.6870
	NET	Model 9	2.9231e+03	0.0479	18.7980
		Model 5	6.7130e+03	0.0874	24.2726

4 and 5, with week 18 to week 33 in 2015 as the test set. A total of 266 influenza samples were included in the high-flu season (weeks 46-15), and the first 250 samples were used as train set to training the models 4 and 5, with week 52 in 2014 to week 15 in 2015 as the test set. The prediction results of the established model are shown in Tables 2-5. Model 6 and model 7 represent the improved influenza prediction models based on models 4 and model 5 using low-flu season data; models 8 and 9 represent improved models 4 and 5 using high-flu season data.

Tables 2-5 indicate that in the low-flu season, model 6 is better than model 4 and model 7 is better than model 5 in both the LS and GA-BP methods. In the high-flu season, model 8 is better than model 4 and model 9 is superior to model 5 in both the LS and GA-BP methods. The experimental results show that the prediction model of seasonal influenza is superior to the models 4 and 5. Because the overall ILI data are divided into high-flu season and low-flu season flu data, making the ILI values of the high-flu season and low-flu season relatively stable with only small changes, it is easy to capture the law and predict future outbreaks.

Prediction-error results for low and high-flu seasons in Region 10 are shown in Fig. 5 and Fig. 6 as an example. Fig. 5, Table 2 and Table 3 show that in the low-flu-season model, the predictive effect of model 6 is better than that of model 4, and model 7 is superior to model 5. Fig. 6, Table 4 and Table 5 show that in the high-flu-season model,

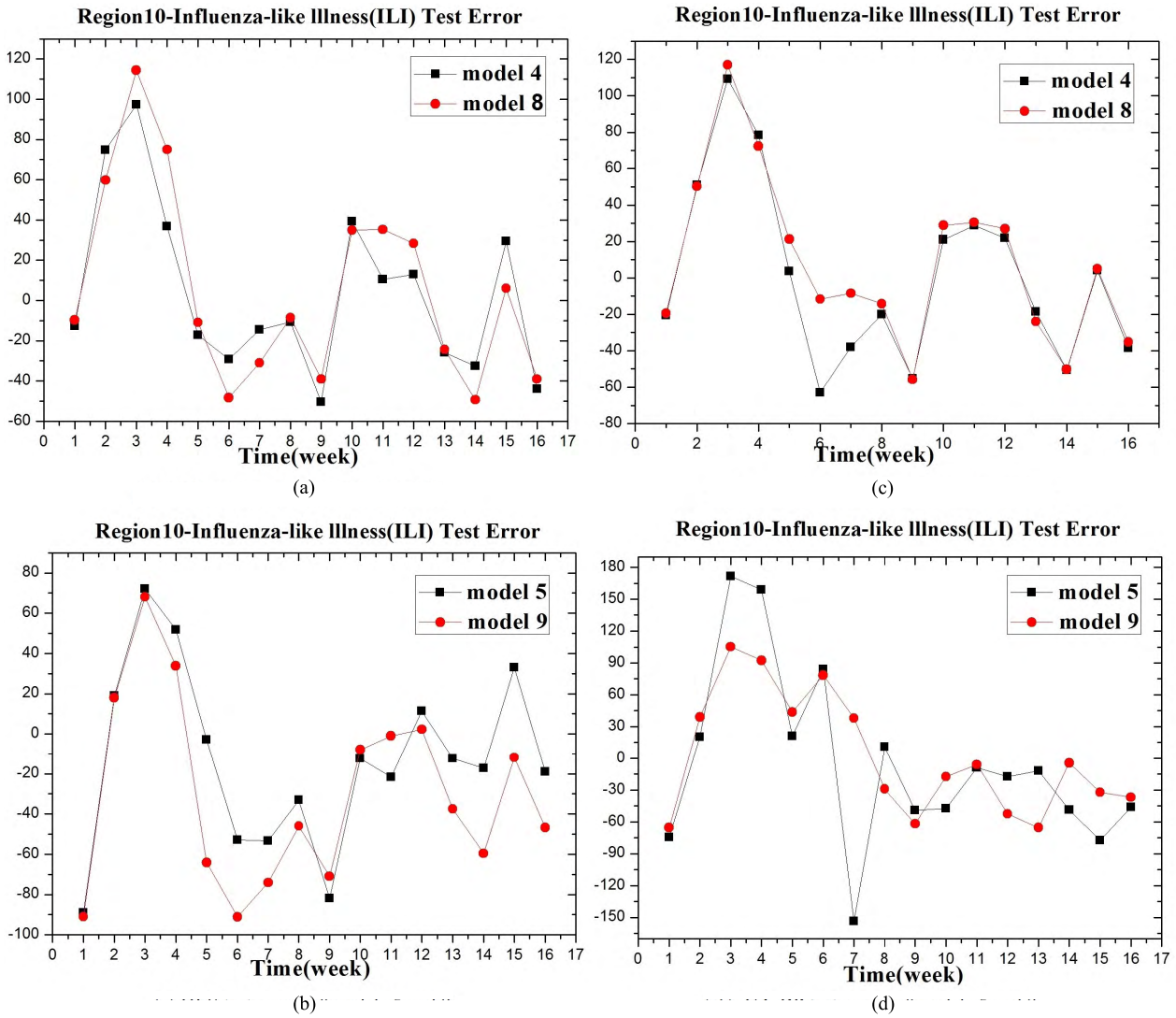


FIGURE 6. Prediction-error chart of Region 10 in the high-flu season. (a) LS test error of models 4 and 8. (b) GA-BP test error of models 4 and 8. (c) LS test error of models 5 and 9. (d) GA-BP test error of models 5 and 9.

the predictive effects of model 8 and model 9 are better than those of model 4 and model 5. Fig. 5 and Fig. 6 and Tables 2-5 show that the prediction results of the GA-BP method are superior to those of the LS method, and the quarter flu-prediction model can be used to monitor and predict flu epidemics to achieve early detection, prevention, and harm reduction.

IV. CONCLUSION

Influenza is a serious - often deadly - and widespread disease that causes substantial economic loss and severe impact to societies across the globe on a yearly basis. Accurately monitoring and tracking influenza activity can help to reduce the spread of the disease and minimize its risks. Public-health emergency-response measures can be determined based on flu monitoring data, and early and efficient flu warning systems can save manpower and expense. In this paper, the regression model for influenza was established with CDC and GFT ILI data. The proposed model was established based on the GFT and CDC regression models of flu outbreaks and the interactions between ten regions of the United States.

The results presented here show that GFT data and historical influenza data are partially complementary in terms of prediction-model performance. The accuracy of real-time prediction can indeed be ensured by the Google search data, while dependent-variable trends are better predicted by historical data. Therefore, the combination prediction model was established by integrating Google search data into the historical data. The results show that the GFT+CDC regression model is preferable for monitoring influenza activity compared to the GFT and CDC regression models. After adding an extra component to the model to include seasonal information (high-flu versus low-flu seasons), the model performed even better.

The theoretical basis for the application of artificial neural networks in the field of regression analysis is discussed here, as a nonlinear regression based on an artificial neural network was utilized for practical analysis. Taking the BP network as an example, a nonlinear regression method based on a GA-BP neural network was proven feasible for the prediction of ILI cases, and the prediction effects were better than LS method.

The models established in this study provide a scientific reference for the future prediction and prevention of influenza. Reducing the risk of influenza via effective monitoring will benefit human health and prevent economic losses across the globe.

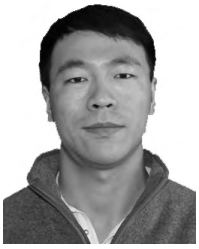
REFERENCES

[1] H. N. Zhai, *The Analysis of Relationship Between Influenza and Atmospheric Ambient and the Establishment of Influenza-Like-Illness Rates Forecasting Model*. Wuhan, China: China Univ. Geosciences, 2009, pp. 13–35.
[2] *National Influenza/Avian Influenza Surveillance Program*, Chin. Nat. Influenza Center, Beijing, China, 2005, pp. 2005–2010.
[3] J. M. Huang, *Detection and Prediction Algorithm for Flu Based on Social Network*. Shanghai, China: East China Normal Univ., 2015, pp. 8–15.
[4] X. T. Li, F. Liu, J. C. Dong, and B. F. Lü, "Detecting China influenza using search engine data," *Syst. Eng.*, vol. 33, no. 12, pp. 3028–3034, Dec. 2013.
[5] J. U. Espino, W. R. Hogan, and M. M. Wagner, "Telephone triage: A timely data source for surveillance of influenza-like diseases," in *Proc. AMIA Annu. Symp.*, 2003, pp. 215–219.
[6] M. Salathé and S. Khandelwal, "Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control," *PLOS Comput. Biol.*, vol. 7, no. 10, p. e1002199, Oct. 2011.
[7] H. A. Johnson *et al.*, "Analysis of Web access logs for surveillance of influenza," in *Proc. 11th World Congr. Med. Inform.*, vol. 107. 2004, pp. 1202–1206.
[8] Z. H. Zhou, H. R. Zhang, and J. Xie, "Data crawler for Sina Weibo based on Python," *J. Comput. Appl.*, vol. 34, no. 11, pp. 3131–3134, Aug. 2014.
[9] S. F. Magruder, "Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease," *Johns Hopkins APL Tech. Dig.*, vol. 24, no. 2, pp. 349–353, Oct. 2003.
[10] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "Twitter: Big data opportunities—Response," *Science*, vol. 345, no. 6193, pp. 148–149, Jul. 2014.
[11] F. Wang *et al.*, "Regional level influenza study with geo-tagged Twitter data," *J. Med. Syst.*, vol. 40, no. 8, p. 189, Aug. 2016.
[12] D. A. Broniatowski, M. Dredze, M. J. Paul, and A. Dugas, "Using social media to perform local influenza surveillance in an inner-city hospital: A retrospective observational study," *JMIR Public Health Surveill.*, vol. 1, no. 1, p. e5, Jan/Jun. 2015.
[13] M. J. Paul, M. Dredze, and D. Broniatowski, "Twitter improves influenza forecasting," *PLoS Currents*, vol. 6, Oct. 2014.
[14] D. A. Broniatowski, M. J. Paul, and M. Dredze, "National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic," *PLoS ONE*, vol. 8, no. 1, p. e83672, Dec. 2013.
[15] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza a H1N1 pandemic," *PLoS ONE*, vol. 6, no. 5, p. e19467, May 2011.
[16] P. Copeland, R. Romano, T. Zhang, G. Hecht, D. Zigmond, and C. Stefansen, "Google disease trends: An update," Int. Soc. Neglected Tropical Diseases, London, U.K., Tech. Rep., 2013.
[17] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009.
[18] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of Google flu: Traps in big data analysis," *Science*, vol. 343, no. 6, pp. 1203–1205, Mar. 2014.
[19] M. W. Davidson, D. A. Haim, and J. M. Radin, "Using networks to combine 'big data' and traditional surveillance to improve influenza predictions," *Sci. Rep.-U.K.*, vol. 5, p. 8154, Jan. 2015.
[20] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi, "Assessing Google flu trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic," *PLoS ONE*, vol. 6, no. 8, p. e23610, Aug. 2011.
[21] A. C. Lowen, S. Mubareka, J. Steel, and P. Palese, "Influenza virus transmission is dependent on relative humidity and temperature," *PLoS Pathogens*, vol. 3, no. 10, p. e151, Oct. 2007.
[22] I. V. Polozov, L. Bezrukov, K. Gawrisch, and J. Zimmerberg, "Progressive ordering with decreasing temperature of the phospholipids of influenza virus," *Nat. Chem. Biol.*, vol. 4, no. 4, pp. 248–255, Mar. 2008.
[23] L. A. Ezzine. (1999). *NIH Scientists Find Why Flu Virus Spreads in Cold Weather*. [Online]. Available: https://www.integrativepractitioner.com/archive/nih-scientists-find-why-flu-virus-spreads-in-cold-weather/
[24] G. Hongshen and T. Youde, "Research on non-linear regression forecast models based on artificial neural networks," *J. North China Univ. Technol.*, vol. 11, no. 1, pp. 68–73, Mar. 1999.
[25] J. Pati, B. Kumar, D. Manjhi, and K. K. Shukla, "A comparison among ARIMA, BP-NN, and MOGA-NN for software clone evolution prediction," *IEEE Access*, vol. 5, pp. 11841–11851, 2017.
[26] X. Zou, W. Zhu, L. Yang, and Y. Shu, "Google flu trends—The initial application of big data in public health," *Chin. J. Preventive Med.*, vol. 49, no. 6, pp. 581–584, Jun. 2015.
[27] O. M. Araz, D. Bentley, and R. L. Muelleman, "Using Google flu trends data in forecasting influenza-likeŰillness related ED visits in Omaha, Nebraska," *Amer. J. Emerg. Med.*, vol. 32, no. 9, pp. 1016–1023, Feb. 2013.
[28] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen, "Reassessing Google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales," *PLoS Comput. Biol.*, vol. 9, no. 10, p. e1003256, Oct. 2013.
[29] J. R. Ortiz, H. Zhou, D. K. Shay, K. M. Neuzil, A. L. Fowlkes, and C. H. Goss, "Monitoring influenza activity in the United States: A comparison of traditional surveillance systems with Google flu trends," *PLoS ONE*, vol. 6, no. 4, p. e18687, Apr. 2011.
[30] *CDC-Seasonal Influenza (Flu)-Flu Activity and Surveillance*. [Online]. Available: http://www.cdc.gov/flu/weekly/
[31] X. C. Wang, F. Shi, L. Yu, and Y. Li, *The Analysis of 43 Cases in MATLAB Neural Network*. Beijing, China: Beijing Univ. Aeronautics Astronautics Press, 2013, pp. 20–33.
[32] S. X. Wang, N. Zhang, L. Wu, and Y. Wang, "Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method," *Renew. Energ.*, vol. 94, pp. 629–636, Aug. 2016.
[33] L. Ke, G. Wenyan, S. Xiaoliu, and T. Zhongfu, "Research on the forecast model of electricity power industry loan based on GA-BP neural network," *Energy Procedia*, vol. 14, pp. 1918–1924, Dec. 2012.
[34] S. Yu, X. Guo, K. Zhu, and J. Du, "A neuro-fuzzy GA-BP method of seismic reservoir fuzzy rules extraction," *Expert Syst. Appl.*, vol. 37, no. 3, pp. 2037–2042, Mar. 2010.
[35] M. B. Bashir and A. Nadeem, "Improved genetic algorithm to reduce mutation testing cost," *IEEE Access*, vol. 5, pp. 3657–3674, 2017.
[36] *Google Flu Trends*. [Online]. Available: http://www.google.org/flutrends
[37] *How Does This Work?* [Online]. Available: https://www.google.org/flutrends/about/how.html

HONGXIN XUE received the M.S. degree from the Department of Mathematics, North University of China, China, in 2015, where she is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering. Her current research interests include artificial intelligence, signal processing, support vector machines, and intelligent optimization algorithm.

YANPING BAI received the Ph.D. degree from the North University of China, China, in 2005. She is currently a Professor and a Doctoral Tutor with the School of Information and Communication Engineer, North University of China. Her research interests include optimization algorithm, artificial intelligence, signal processing, image processing, and MEMS reliability research.

HONGPING HU received the Ph.D. degree from the North University of China, China, in 2009. She is currently a Professor and a Master Tutor with the Department of Mathematics, North University of China. Her research interests include combinatorial mathematics, artificial intelligence, and image processing.



HAIJIAN LIANG received the M.S. degree from the School of Information and Communication Engineering, North University of China, China, in 2015, where he is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering. His current research interests include temperature measurement and signal processing.