Journal of Computers Vol. 31 No. 2, 2020, pp. 101-113
doi:10.3966/199115992020043102010

Research on the Public Opinion Early Trend Prediction Based on the Trend Similarity of Emergency

Fu-Lian Yin[1,] Xiao-Wei Liu[1], Bei-Bei Zhang[1*]

[1] College of Information and Communication Engineering, Communication University of China, Beijing, China
yinfulian@cuc.edu.cn, cuclxw@163.com, cuc_zhangbeibei@qq.com

Received 13 June 2018; Revised 17 October 2018; Accepted 18 December 2018

Abstract. Under the environment of micro-blog communication in China, the current public opinion trend forecasting models can only forecast the trend of a complete single event, but the result of the early trend forecasting is very poor. To solve this problem, an early trend predicting algorithm based on the develop trend similarity of hot events is put forward in this paper. We apply the clustering results of the time series of hot events to build the prediction model for different categories, in which the GABP (genetic algorithm optimized BP neural network) is used to build the prediction model. We also propose the matching rules between a new event and the existing category. According to these, we realize the developing trend prediction for the new events. The experiment shows that the early trend forecasting algorithm proposed in this paper has better accuracy and timeliness when predicting the early development trend of public opinion. The APE (Absolute Percentage Error) of about 75% samples is below 80%, MSE (Mean Square Error) is below 0.01. And comparing with the traditional predicting algorithm, MSE is decreased by 90%, and APE decreased by 24%.

Keywords: BP neural network, classified prediction, combination prediction, genetic algorithm, time series matching

1 Introduction

Micro-blog covers the frontier internet public opinion of China. When hot issues erupt, discussions on this platform often represent the most direct voice of Chinese netizens. Therefore, it is very important to grasp the development trend of micro-blog public opinion. However, the development of hot events is often difficult to predict, because of the large impact on society, complexity, high sensitivity and serious consequence. The improper solutions and methods will cause a lot of adverse social reactions and harm social stability. Most of predicting algorithms of public opinion predict the future development according to the historical data, which isn't able to forecast the trend of the event just when it occurs. Although the long-term trend prediction of the topic is significant, the early trend prediction of the network public opinion is also very important, because the life cycle of topics is getting shorter and shorter. In fact, as for many widely concerned public opinion events, there are only a few days between the appearance and eruption.

At present, the researches mainly focus on the long-term prediction of the topics, including the prediction model based on the traditional statistics and the machine learning [1]. The prediction models based on the traditional statistical include Logistic model [2], exponential smoothing model [3], ARIMA model [4] and moving average model [5] and so on. These kinds of model require that the forecasting data has a strong regularity in order to get a better fitting effect. The prediction models based on machine learning algorithm mainly combine artificial intelligence technology and time series prediction, including neural networks [6-10], gray theory [11-12], Bayesian network [13-14], fuzzy set theory [15-16] and so on. These prediction models have a strong ability to predict fuzzy prediction problem and nonlinear

[*] Corresponding Author

prediction problem. The advantages and disadvantages of the various prediction models are as Table 1.

Table 1. Advantages and disadvantages of various forecasting models

Forecasting methods	Advantages	Disadvantages
Logistic regression model	The data set is small and the calculation is simple, more suitable for the short-term forecast.	Can only study the monotonous curve and the data processing ability is poor.
Exponential smoothing model	It introduces the weighted average method and supports to set the weights of historical data.	Difficult to determine the exponential smoothing coefficient, influenced subjectively.
ARIMA model	The model is simple and the calculation is simple and it's suitable for the periodic prediction.	Can only capture the linear relationship essentially.
Moving average model	Can reveal the long-term trend of time series.	The forecasting value always stays at the past level and can't predict the higher or lower fluctuations in the future.
BP neural network predicting model	It has a wide range of adaptability, learning ability and mapping capabilities.	Convergence speed is slow, and it's easy to fall into the local minimum.
Gray theory	Suitable for the prediction of some data that grow exponentially.	The prediction results of the time series with big wave properties is poor.
Bayesian Network	It fits the data with incomplete evidence. It can make full use of the priori information to make probabilistic reasoning prediction.	Can only predict the warning level of public opinion development, can't predict the future development trend of public opinion accurately.
Fuzzy set theory	Its mathematical description is more in line with the nature of the fuzzy object in the objective world.	Can only predict the warning level of public opinion development, can't predict the future development trend of public opinion accurately.

In the above forecasting methods, the data in early stage of a single event is always set as the training data, the others as the test data for model training. As we all know, the development of public opinion is generally unstable in the early stage, relatively stable in the late stage. Therefore, testing a forecasting model based on the relatively stable data, the expandability and universality of the model can't be guaranteed. And the forecasting model using the pre-data as training data can only predict the development trend in the middle and late period more accurately, which can't predict the changes when the event just occurs, leading to the awful timeliness of the model. To solve this problem, we find neural network has well ability to deal with nonlinear mapping problems. At the same time, neural network has the characteristics of self-learning, self-adapting and good flexibility, so it is more convenient to carry out. In addition, the genetic algorithm is a method to simulate the natural evolutionary process and search for the optimal solution of the problem. Its main characteristics are that it operates directly on structural objects without restriction of derivative and function continuity, and has inherent implicit parallelism and better global optimization ability. Several scholars use genetic algorithm to improve BP neural network in different fields have achieved good results, greatly improving the convergence speed and prediction accuracy of BP neural network [10, 17]. Based on the comparison with previous researches, using GABP to achieve a better performance in prediction is worth researching.

As for the appliance of prediction, there are abundant researches in many fields, such like product and stock [18-19]. And in the field of public opinion, researches about trend prediction based on neural network are widely studied recently [20-22]. Other method such as collaborative filtering and parameter inversion are also proposed and perform well [23-24]. However, the relative research about using GABP to realize the trend prediction is seldom seen. In addition, rare research focuses on early trend prediction of public opinion, which is worth studying for its practical application value. Therefore, the main goal to be achieved in this paper is to make good use of GABP and propose a well-performed scheme to carry out public opinion early trend prediction.

In this paper, based on GABP, we propose an early trend forecasting method based on the trend similarity of hot events. Firstly, the trend prediction scheme in early stage of hot events is proposed in section 2. Then in section 3, the matching rule between the new event and the existing categories is proposed, we use time series cluster algorithm to get different public opinion development trends. And in section 4, neural network prediction model optimized by genetic algorithm is proposed to carry out the prediction. We use the similarity between the events to train the neural network model repeatedly to make the prediction model more robust and applicable. When the new events occur, we can match the new data with the existing category, so we can use the existing forecasting model for public opinion trends prediction. In section 5, Experiments and result analysis show that the proposed algorithm possesses high accuracy when predicting the new events. Finally, Section 6 concludes this paper.

2 The Trend Prediction Scheme in Early Stage of Hot Events

In this paper we present a new forecasting scheme for public opinion. Firstly, the matching rules of the new event and the existing categories are established, and the data of the new event is processed and matched with the existing C categories. When getting the class C' with the highest matching degree, take the class C' as the category of the new event. Then the public opinion data in the class C' are processed to input the neural network prediction model, and the time series prediction model of class C' is established with the BP neural network prediction algorithm optimized by the genetic algorithm. Finally, the data of new event is input into the prediction model of class C', and the new event public opinion time series is predicted (see Fig. 1).

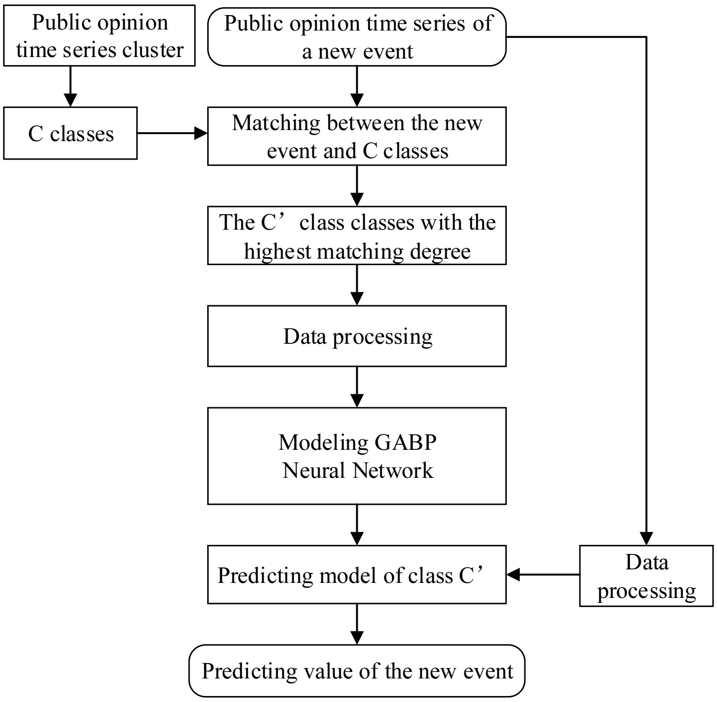
Public opinion time series cluster
Public opinion time series of a new event
C classes
Matching between the new event and C classes
The C' class classes with the highest matching degree
Data processing
Modeling GABP Neural Network
Predicting model of class C'
Data processing
Predicting value of the new event

Fig. 1. Circuit for computing $P'(x)$

Step1. Time series clustering. According to the hierarchical clustering method based on S-DTW (Segmented Dynamic Time Warping distance), the existing events in the public opinion time series database are clustered as several classes with different development trend. The hierarchical clustering method based on S-DTW is the leading work of this paper. S-DTW (Dynamic Time Warping distance) is a method to calculate the similarity of time series proposed by the writer. This method can make the time series information loss less when clustering the time series, the computational complexity is greatly reduced comparing with the traditional DTW calculation method and the contour coefficient is also improved. As a result, we can get a well-functioning clustering effect.

Step2. The matching rule between new event and the existing category. Here, we propose a new method of matching the new event with the existing category when the new event happens. After matching the existing category according to this method, the best matched class is regarded as what the new event belongs to.

Step3. Training neural network prediction model. When getting the best matched class, we train the neural network prediction model for the class and GA (genetic algorithm) is introduced into the training of traditional BP neural network, the initial parameters of BP neural network are optimized so the prediction accuracy of neural network is improved.

The two key steps of this scheme lie in step 2 and 3. Step 2 ensures that the new event matches the category exactly what it belongs to in the existing ones. Step 3 ensures that the neural network prediction model of a one category has far better prediction ability and high prediction accuracy compared with models of other classes. These two conditions are the key factors in the establishment of the scheme, only if the new event matches its belonging category, and each category has strong targeted prediction ability, it is possible to make accurate predictions for the new event. These two conditions are verified in the module 3 and 4, which proves the feasibility of this scheme.

3 The Matching Rule between the New Event and the Existing Categories

In this paper, a matching method between new events and existing categories is established. When a new event is detected, construct the complete public opinion time series for predicting the development trend of new events, on which we can calculate the similarity between the new event and each sample in each category. Finally, get the average similarity between the new event and each category, and then keep the c (c <= C) categories as the matched clusters, C is the total number of categories.

3.1 Construct the New Public Opinion Time Series

When a new event is detected, the opinion heat of the event is collected every 4 hours x_{nj}, finally get the public opinion heat time series $x_{11}, \cdots, x_{1l_{st}}, \cdots, x_{n1}, \cdots, x_{nj}$. We define the length of the sequence as the real length of the sample. n is the number of days from the event occurrence day to the collecting day, j represents the j_{th} collection point in one day of the event, l_{st} represents the real sample length of the first day after the event's occurrence, $n \geq 1, 1 \leq j, l_{st} \leq 6$. We fill several 0 before the sample data in the first day of the event to indicate somewhere the occurrence time point in the day. In addition, we define the length of the new sequence as the valid length of the sample. The construction method of new public opinion time series is as Fig. 2.

```
Begin
    The data will be divided by natural day,
    Get $\{x_{11}, \cdots, x_{1l_{st}}\}, \cdots, \{x_{n1}, \cdots, x_{nj}\}$,
    Calculate the real sample length of the first day of the event $l_{st}$,
    Fill $6 - l_{st}$ 0 before the sample data of the first day of the event,
    Get new public opinion heat time series $x = \{rep(6 - l_{st}), x_{11}, \cdots, x_{1l_{st}}\}, \cdots, \{x_{n1}, \cdots, x_{nj}\}$,
End
```

Fig. 2. Construction method of new public opinion time series

3.2 Sequence Similarity Calculation Method Based on Variable Length Window

In this paper, a similarity calculation method based on variable-length window is designed to calculate the similarity between the new event and each sample within the category. The calculation method is shown in Fig. 3. The existing subsequence matching algorithms mostly use DTW distance, and they are realized traversing the time series by the sliding window. The algorithms have a high time complexity. When calculating the similarity of time series, in order to make full use of the initial data of the new event, the starting point of the time series of new event should be aligned with the data of the existing sample. What's more, in order to ensure that the new event and the existing sample reach a minimum distance within a certain range, this paper proposes the following method. We set up a scalable window whose length is adjusted near the length of the existing sample time series. By this way, it is possible to calculate the similarity between the new event and the sample in the existing database and further calculate the similarity between the new event and the existing category. The algorithm has a lower computational complexity.

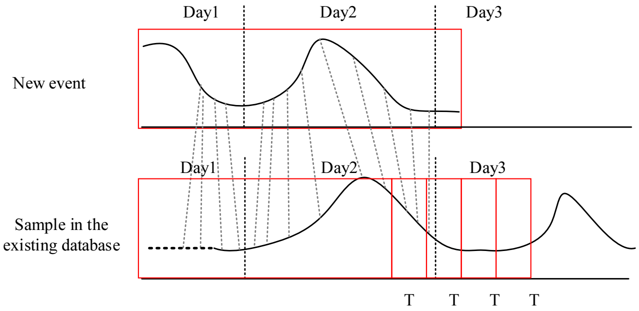

Fig. 3. Sub-sequence similarity calculation based on the variable length window

Set the time series of new event as the target sequence, and the other existing sample as the matching sequence. If the length of the target sequence is T_0, the length of window is $T_0 - mT$ to $T_0 + mT$, and then calculate the time series distance of the target sequence and the matching sequence in these windows. The minimum value is the final distance of the target sequence and the matching sequence. The time series distance calculating method is based on the S-DTW distance method in paper [25].

Algorithm 1. Calculations of the S-DTW distance between time series *x* and *y*

```
Given:
$\boldsymbol{x} = \{x_{11}, ..., x_{1a}, x_{21}, ..., x_{26}, ..., ..., x_{n1}, ..., x_{nc}\}$, the first time series with $n$ days
$\boldsymbol{y} = \{y_{11}, ..., y_{1b}, y_{21}, ..., y_{26}, ..., ..., y_{n1}, ..., y_{nd}\}$, the second time series with $m$ days
$1 \le a, b, c, d \le 6$
Output:
$D_{stdw}(\boldsymbol{x}, \boldsymbol{y})$
SegmentByDay($\boldsymbol{x}$);
Get $\boldsymbol{x} = \{\boldsymbol{x_1} = \{x_{11}, ..., x_{1a}\}, \boldsymbol{x_2} = \{x_{21}, ..., x_{26}\}, ..., \boldsymbol{x_n} = \{x_{n1}, ..., x_{nc}\}\}$,
SegmentByDay($\boldsymbol{y}$);
Get $\boldsymbol{y} = \{\boldsymbol{y_1} = \{y_{11}, ..., y_{1b}\}, \boldsymbol{y_2} = \{y_{21}, ..., y_{26}\}, ..., \boldsymbol{y_m} = \{y_{m1}, ..., y_{md}\}\}$
if $a \ge b$ do
    $\boldsymbol{y_1} = \{\text{repete}(0, 6-b), \boldsymbol{y_1}\}$
  else do
    $\boldsymbol{x_1} = \{\text{repete}(0, 6-a), \boldsymbol{x_1}\}$
if $c \ge d$ do
    $\boldsymbol{y_m} = \{\text{repete}(0, 6-d), \boldsymbol{y_m}\}$
  else do
    $\boldsymbol{x_n} = \{\text{repete}(0, 6-c), \boldsymbol{x_n}\}$
if $n \ge m$ do
  dist1 = $\sum_{i=1}^{m} D_{tw}(\boldsymbol{x_i}, \boldsymbol{y_i})$
  dist2 = $\sum_{i=m+1}^{n} \boldsymbol{x_i}^2$
$D_{stdw}(\boldsymbol{x}, \boldsymbol{y}) = \text{sqrt}(\text{sum}(\text{dist1}, \text{dist2}))$
if $n < m$ do
  dist1 = $\sum_{i=1}^{n} D_{tw}(\boldsymbol{x_i}, \boldsymbol{y_i})$
  dist2 = $\sum_{i=n+1}^{n} \boldsymbol{y_i}^2$
$D_{stdw}(\boldsymbol{x}, \boldsymbol{y}) = \text{sqrt}(\text{sum}(\text{dist1}, \text{dist2}))$
Return $D_{stdw}(\boldsymbol{x}, \boldsymbol{y})$
```

3.3 Verification of the Matching Method

In this paper, the following experiments are carried out to verify the accuracy of the matching method proposed here. We selected the time series in initial stage of each event with different lengths to match the existing categories, and get three categories with the highest matching degree at different valid sample lengths for each sample. Compare the sample number where the first 3 categories are separately equal to the true category the sample belongs to. The results are shown in the Fig. 4.

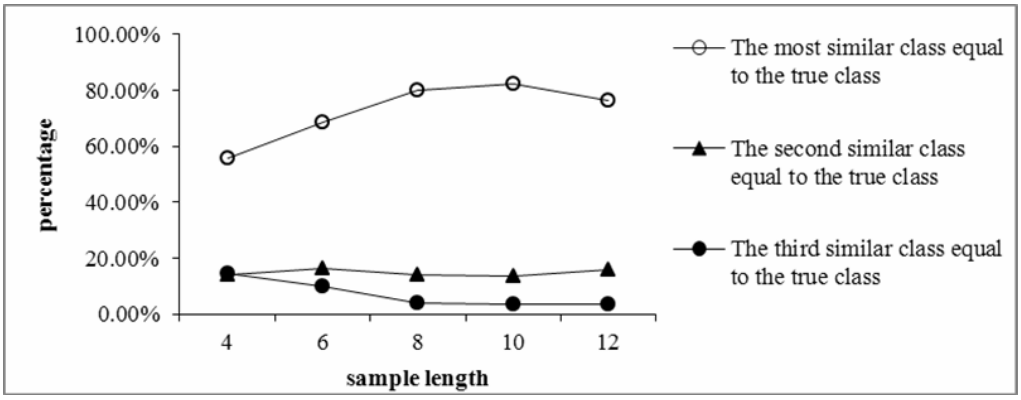

Fig. 4. The matching situation schematic diagram

The above figure proves that the matching method of new events and existing categories in this paper can ensure that about 70% of the samples can match to their belonging categories in the most similar class. Moreover, with the increase of the sample valid length, the ratio keeps rising generally.

4 Neural Network Prediction Model Optimized by Genetic Algorithm

As the development trend of events in each cluster are highly similar, we propose a neural network prediction model for each category in order to make full use of the similarity between classes and predict the trend of new events. The neural network prediction models of different cluster have better robustness and targeted predictive ability because they use the similar trends in each cluster to train one prediction model. This view is validated in the final experiment in this section. In this paper, the traditional BP neural network is used to train the public opinion development trend prediction model, and the genetic algorithm is used to optimize the initial parameters of the neural network, which greatly improves the prediction accuracy of BP neural network.

4.1 BP Neural Network Algorithm

The process of training BP neural network is divided into two stages, the first stage is the forward propagation of the signal, from the input layer through the hidden layer, and finally to the output layer; the second stage is the reverse propagation of the error, from the output layer through the hidden layer, and finally to the input layer, through which adjusting the weight threshold from the hidden layer to the output layer, from the input layer to the hidden layer.

Set the number of samples is *N*, and the input vector of each sample is $X=[x_1,x_2,...,x_n]$, the corresponding expected output vector is $Y=[y_1,y_2,...,y_m]$. The BP neural network training process includes the following steps:

(1) Network initialization.

Determine the number of nodes *n*, *l*, *m* of input layer, hidden layer and output layer in the network, the connection weights between the input layer and the hidden layer ω_{ij}, the connection weights between the hidden layer and the output layer ω_{jk}, the hidden layer threshold $a=[a_1,a_2,...,a_l]$, and the output layer threshold $b=[b_1,b_2,...,b_m]$.

(2) Input the first sample [X_1, Y_1].

(3) Calculate the hidden layer output h_j.

$$h_j = f(\sum_{i=1}^{n} \omega_{ij} x_i - a_j), i = 1, 2, \cdots, n; j = 1, 2, \cdots, l. \quad \textbf{(1)}$$

f is excitation function $f(x) = \frac{1}{1+e^{-bx}}, b > 0$ in the hidden layer, and x_i is the *i_{th}* variable of input node.

(4) Calculate the output layer output o_k.

$$o_k = f(\sum_{j=1}^{l} \omega_{jk} h_j - b_k), k = 1, 2, \cdots, m.. \quad \textbf{(2)}$$

(5) Calculate the network prediction error *e*.

$$e_k = y_k - o_k.. \quad \textbf{(3)}$$

In this function, y_k is the expected network output value.

(6) Update the weights and thresholds according to the network error *e*.

$$\omega_{ij}(t+1) = \omega_{ij}(t) + ah_j(1-h_j)x_i. \quad \textbf{(4)}$$

$$\omega_{jk}(t+1) = \omega_{jk}(t) + ah_j e_k. \quad \textbf{(5)}$$

In this function, $i = 1, 2, \cdots, n; j = 1, 2, \cdots, l; k = 1, 2, \cdots, m.$ *a* is the learning velocity.

$$a_j(t+1) = a_j(t) + ah_j(1-h_j). \quad \textbf{(6)}$$

$$b_k(t+1) = b_k(t) + e_k. \quad \textbf{(7)}$$

(7) Train the next sample and cycle steps (3) - (6) until all samples are finished training.

(8) All the samples are trained in second rounds, and the steps (3) - (6) are carried out until the algorithm exceeds the iteration number or the global error of network is less than the objective error.

The objective function is the mean square error MSE:

$$E = \frac{1}{2N} \sum_{i=1}^{N} \sum_{k=1}^{m} (\hat{y}_{ik} - y_{ik})^2. \quad \textbf{(8)}$$

On the one hand, BP neural network algorithm is a gradient descent method essentially, which brings about that the slow convergence rate of the algorithm. On the other hand, BP neural network is an optimization method for local search, so the weights adjust gradually along the direction of local improvement, which will make the weights converge to the local minimum. In addition, the BP neural network is very sensitive to the initial network weight. It tends to converge to different local minimum with different initial weights. In this paper, genetic algorithm is introduced to determine the initial weights of BP neural network to improve the convergence rate and avoid falling into the local minimum.

4.2 Genetic Algorithm

Genetic algorithm comes from the natural selection and genetics of Darwin. Through selecting, crossing and mutating individuals, it enables the individuals to have the ability of searching best solution globally. The genetic algorithm is introduced to determine the initial weight of BP neural network, raising the convergence rate of the algorithm, preventing it from falling into the local minimum. The BP neural network optimized by genetic algorithm is shown in Fig. 5.

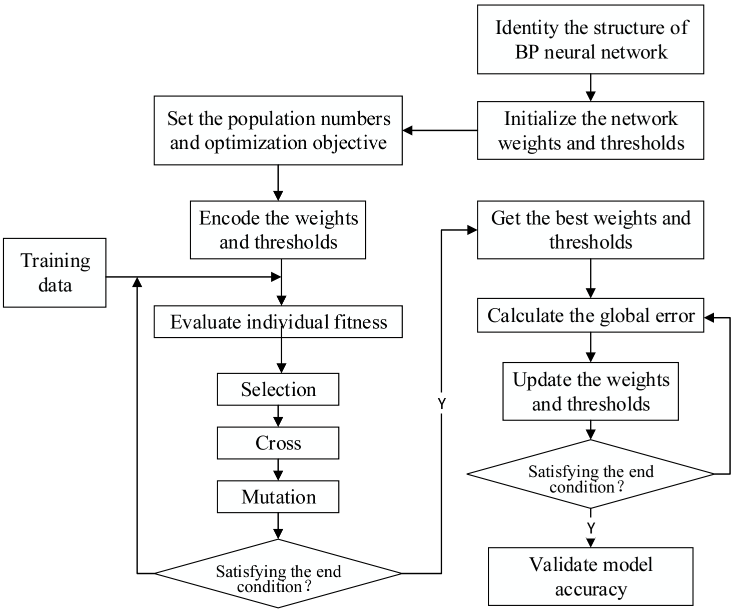
Identity the structure of BP neural network
Initialize the network weights and thresholds
Set the population numbers and optimization objective
Encode the weights and thresholds
Training data
Evaluate individual fitness
Selection
Cross
Mutation
Satisfying the end condition ?
Y
Get the best weights and thresholds
Calculate the global error
Update the weights and thresholds
Satisfying the end condition ?
Y
Validate model accuracy

Fig. 5. BP neural network algorithm optimized by genetic algorithm

(1) Set the population numbers is *P*. Create *P* initial populations as $G=(G_1,G_2,\cdots,G_p)^T$. Select *S* real numbers lying in $[-W,W](W<=3)$ and constitute a vector with length S as the chromosome of each individual G_i.

$$S=n*l+l*m+l+m. \quad \textbf{(9)}$$

n, *l*, *m* is separately the number of nodes in input layer, hidden layer and output layer in the neutral network. Each individual of population $G_i=(g_1,g_2,\dots,g_s),i=1,2,\dots,P$, represents a group of initial value of a neutral network. Each gene g_j of individual G_i represents a weight or threshold.

(2) Calculate the individual fitness. Establish a neutral network according to the best chromosome, train the network until it satisfies the end condition and we can get a group of output. The $fitness_i$ and average fitness $\hat{f}$ of individual G_i in population *G* is：

$$fitness_i=MES. \quad \textbf{(10)}$$

$$f=\frac{\sum_{i=1}^{P}fitness_i}{P},i=1,2,\dots,P \quad \textbf{(11)}$$

(3) Select the chromosomes in each generation population according to “roulette wheel”.

(4) Cross the chromosomes with one-point crossover. Assume that chromosome G_u and G_v cross at the point r_pick, then the two later generations are:

$$\begin{cases}G_u'=G_u[1:r\_pick]+G_v[(r\_pick+1):S].\\G_v'=G_v[1:r\_pick]+G_u[(r\_pick+1):S].\end{cases} \quad \textbf{(12)}$$

r_pick is a random integer in [0, *S*]. *S* is the length of chromosome.

(5) Mutation. Select and mutate the j_{th} gene g_j of *a* individual.

$$g_j=\begin{cases}g_j+(g_{\max}-g_j)f(g), & r>0.5\\g_j-(g_j-g_{\min})f(g), & r\le 0.5\end{cases}. \quad \textbf{(13)}$$

$$f(g)=r(1-\frac{iter_{now}}{iter_{\max}}). \quad \textbf{(14)}$$

$g_{\max}$ and $g_{\min}$ is separately the upper and lower bound of gene g_j. r is a random in [0, 1]. $iter_{now}$ is the current iteration and $iter_{\max}$ is the max evolution iteration.

(6) Allocate the optimal individual of genetic algorithm into the weights and thresholds of BP neural network, and train the BP neural network model with BP algorithm to get the global network error.

(7) Judge whether the algorithm is finished, which means whether the global error is less than the target error or whether the evolution iteration reaches the maximum evolution iteration. If not, return step (3).

4.3 Neural Network Prediction Experiment and Analysis

This experiment is aimed to compare the performances of the optimized genetic algorithm neural network prediction model with traditional neural network prediction model, and verify that after clustering the public opinion time series, the prediction model for each cluster has far better targeted predictive ability.

(1) Evaluation method. In this paper, MSE and MAPE (Mean Absolute Percent Error) and APE are used as evaluation indexes to evaluate the performance of different prediction models. The formula is as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - o_i)^2. \quad \textbf{(15)}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}|(y_i - o_i)/o_i| \times 100\%. \quad \textbf{(16)}$$

$$APE = |(y_i - o_i)/o_i| \times 100\% \quad \textbf{(17)}$$

y_i is the actual value, o_i is the predicted value. N is the total number of samples.

(2) Data preparation for neural network prediction model. After clustering the time series of hot events, we get the different development trends of various clusters as shown below.

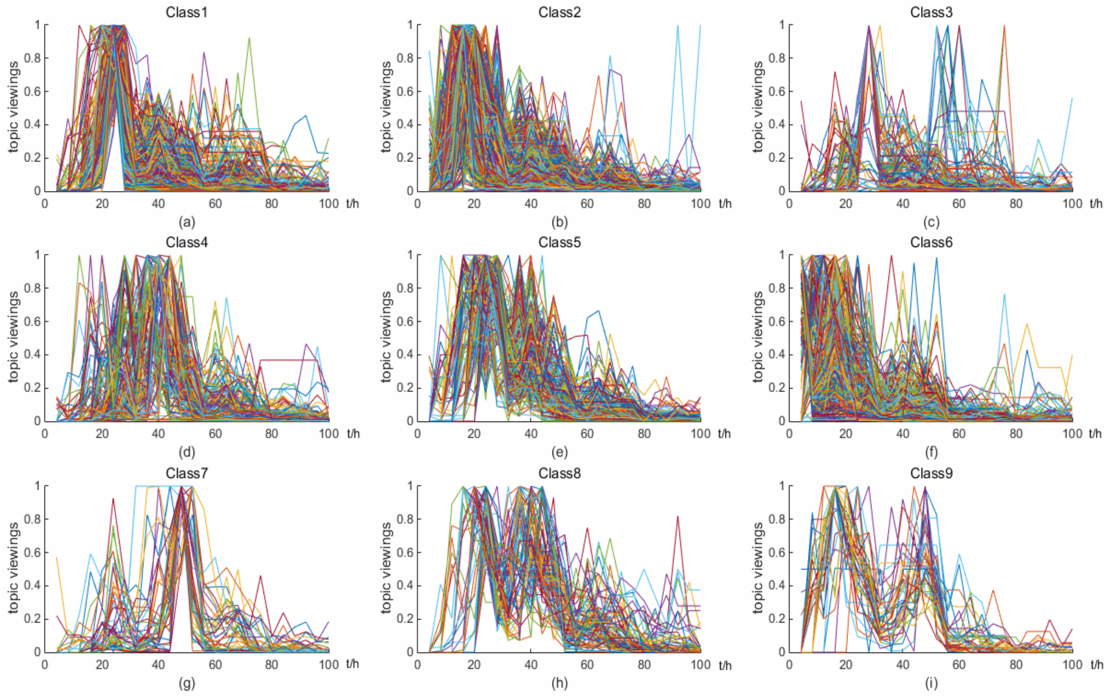

Fig. 6. The public opinion trends of different clusters for hot events

Since the time series used here are from hot events and the development rule of hot events is usually cycled by 24 hours, the data sampling period is 4 hours in this paper. So there are six viewing data in one day, the number of neurons in input layer in the neural network can be set as 6, the number of neurons in the output layer as 1, and the number of neurons in the hidden layer as 10. Therefore, the structure of the neural network used here is 6-10-1. Then all kinds of cluster data are organized into 6-1 structure circularly.

(3) Performances comparison between GABP and BP. The data of the 9 categories are all divided into training set and test set. The training set take up 90% of sample number of each cluster. The BP neural network and GABP neural network are trained on the same training set, and tested on the same test set. The experiment parameters are as follows:

BP neural network algorithm. BP neural network with 3 layers: the input layer 6 nodes, hidden layer 10 nodes, output layer 1 node. The transfer function of hidden layer is S function named tansig, the output layer is the S function named logsig. The training algorithm uses levenberg-marquart algorithm. The maximum number of training is set as 1000, training objective error as 0.00001, the learning rate as 0.1.

Genetic algorithm. The population size is set as 20, evolution iteration as 50, crossover rate as 0.8 and the mutation rate as 0.1.

Table 2 shows that the performance of GABP neural network is far beyond the traditional BP neural network. Therefore, the prediction models of the other experiments in this paper adopt GABP neural network prediction model completely.

Table 2. Performances comparison between GABP neural network and traditional BP neural network

Class	MSE		MAPE	
	BP	GABP	BP	GABP
Class1	0.2865	0.0065	2.9177	0.6593
Class2	1.3612	0.0042	1.8807	0.7006
Class3	1.4844	0.0262	2.0534	0.9802
Class4	1.5440	0.0350	1.6780	0.8239
Class5	0.5405	0.0166	2.3841	0.6620
Class6	0.0829	0.0019	2.2747	0.7607
Class7	0.1081	0.0359	1.3509	0.9488
Class8	0.2847	0.0179	1.3490	0.4901
Class9	0.2322	0.0356	1.8414	1.0144

5 Experiment and Results Analysis

5.1 Predictive Effectiveness Evaluation of Early Trend Forecasting Method

In this experiment, the valid sample length is set as 4, 6, 8, 10, 12, corresponding to the 16, 24, 32, 48 hours after the event occurs. Compare the predictive effect of the early trend forecasting methods proposed in this paper when choosing different valid sample length. The distribution of the prediction error MSE is shown in Fig. 7.

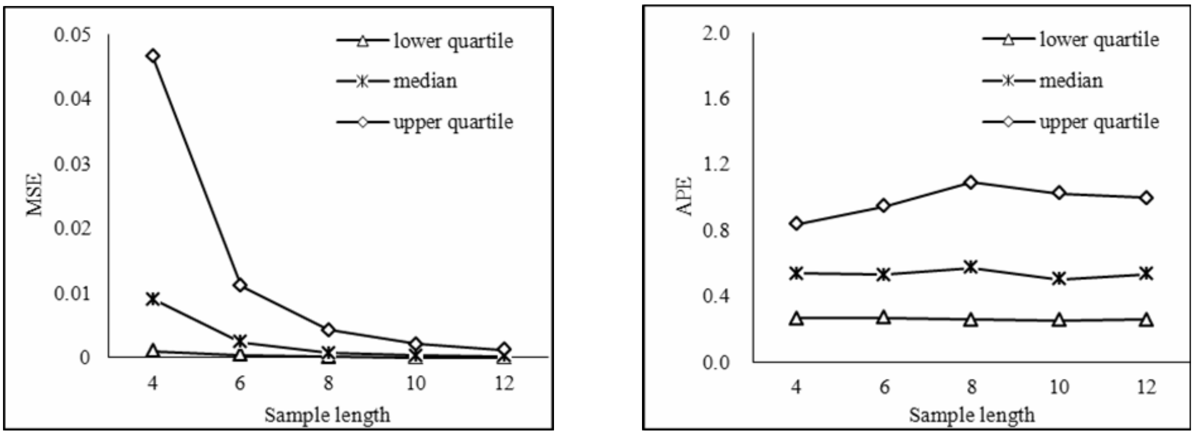

Fig. 7. Distribution of MSE and APE with different valid sample length

The graph shows that MSE is closer to zero as the length of the sample increases generally, indicating that the prediction error tends to decrease. APE of about 75% samples is below 80%, APE of about 25% samples is between 80% and 100%. Combined with Fig. 7, it shows that with the increase of the sample length, the probability of the sample matching to its belonging category is growing, the prediction error is getting smaller and smaller, which means, the predictive effect of the public opinion early trend prediction scheme is getting better with the increase of the sample length.

5.2 Effectiveness Comparison of Early Trend Forecasting Method and Traditional Forecasting Method

The most important feature of public opinion forecasting method proposed in this paper is forecasting new events based on the similarity of the trends of hot event. It is necessary to establish a set of hot events with different development trends. The traditional public opinion trend prediction method establishes the forecasting model based on the historical data of a single event and predict the later development trend. Here we compare the forecasting effects of early trend forecasting methods of public opinion with traditional forecasting methods such as ARIMA, gray prediction, BP neural network and so on. In the experiment, we set the valid sample length as 4, which means 16 hours after the event occurrence. The prediction error MSE distribution is shown in Fig. 8.

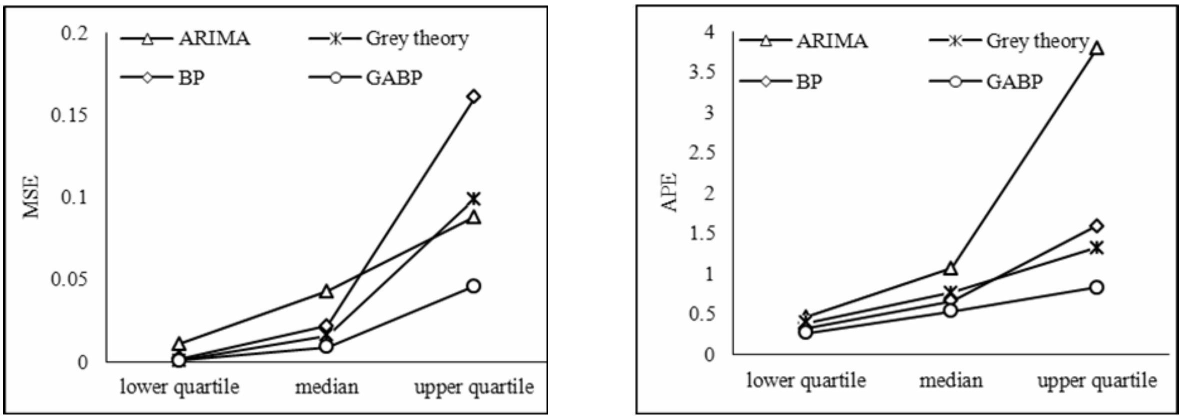

Fig. 8. MSE, APE distribution of different prediction methods

Fig. 8 shows that the MSE and APE distribution of the early trend forecasting method based on the similarity of hot events are much lower than those of other traditional prediction methods. The MSE of this method is reduced by at least 90% compared with the traditional prediction method, and the APE is reduced by at least 24%, indicating that the proposed method of early trend prediction of public opinion has high predictive accuracy and timeliness.

6 Summary

Nowadays, the current public opinion trend forecasting models can only forecast the trend of a complete single event, but the results of the early trend forecasting is very poor. Therefore, we propose an early trend forecasting method based on the similarity of hot events under the environment of micro-blog communication in China. Firstly, we put forward an assumption that establishing a forecasting model for different categories with different trends can predict the development trend of hot events more accurately. The hypothesis is proved through some experiment. Then, we propose the matching rule between the new event and the existing categories. By carrying out some experiments we prove that matching rule can make a new event matches to its belonging class when the event just occurs with a high accuracy, ensuring the predictive accuracy of the prediction models for each class. Finally, a combination forecasting method is proposed, by which we get the final prediction value. The experiment proves that the MSE and APE of the combination forecasting method is about 4%-35% lower than the general weighting method. We carry out experiment with the public opinion early trend forecasting method to evaluate the prediction effect. We found that the predictive accuracy of this method is very high when the new event just occurs, the APE of about 75% of the sample is below 80% stably, and about 25% is between 80%-100%. Besides, with the increase of the sample length, the probability of matching the samples to their respective categories is increasing and the prediction error is getting lower and lower. In addition, this paper also compares the forecasting method proposed in this paper with traditional methods such as ARIMA, gray prediction and BP neural network. It is found that the MSE of this method is reduced by at least 90% and APE by at least 24% compared with the traditional prediction methods. In summary, the method proposed in this paper has high predictive accuracy and timeliness for public opinion early trend forecasting. And derivative public opinion is a relatively new research point in micro-blog research, which is the focus of our future research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61801441) and the Fundamental Research Funds for the Central Universities (CUC8Q45).

References

[1] D.-D. You, F.-J. Chen, The literature review about the prediction of network public Opinion in China, Information Science of China 34(12)(2016) 156-160.

[2] Y.-X, Lan, R.-X, Zeng, Research on network public opinion propagation and early warning stage of emergencies, Journal of Information of China 32(5)(2013) 17-19.

[3] M.-J. Xu, Research on microblogging public opinion forecast model based on exponential smoothing, China Public Security 27(42)(2016) 80-84.

[4] R.N. Calheiros, E. Masoumi, R. Ranjan, R Buyya, Workload prediction using ARIMA model and its impact on cloud applications' QoS, IEEE Transactions on Cloud Computing 3(4)(2014) 449-458.

[5] S. Cao, Y.-X. Lan, G.-Q. Su, Research on micro-blog public opinion prediction model based on moving average method, Journal of Hubei University of Police 27(3)(2014) 40-42.

[6] Y.-X. He, J.-B. Liu, N. Liu, Q. Chen, J. He, Topic trend prediction of micro-blog based on improved population model, Journal on communication 29(5)(2015) 5-12.

[7] Y. Shu, L.-L. Zhang, Prediction of network public opinion based on wavelet analysis and artificial neural network, Information science of China 34(4)(2016) 40-42.

[8] N.-N. Wang, W.-J. Zhang, L. Niu, Sentiment prediction of public sentiment based on ARIMA and BP neural network model, Electronics technology 29(5)(2016) 83-87.

[9] D.-D. You, F.-J. Chen, Research on prediction of network public opinion based on improved particle swarm optimization and BP neural network, Journal of Information of China 35(8)(2016) 156-161

[10] S.-K, Zhang, J.-C. Lv, X.-S. Yuan, S. Yin, BP neural network with genetic algorithm optimization for prediction of geo-stress state from wellbore pressures, International Journal of Computational Intelligence & Applications 15(3)(2016) 80-85.

[11] W.-J. Li, C.-C. Hua, W.-Q. He, Grey prediction model and case analysis of Internet public opinion events, Information science of China 31(12)(2013) 51-56.

[12] Z.-T. Du, X.-Z. Xie, Prediction and early warning model of network public opinion based on grey prediction and pattern recognition, Library and Information Service 57(15)(2013) 27-33.

[13] J.-L. Fang, J.-M. Huang, M. Liu, The research on fuzzy Bayesian network model for the network public opinion situation and threat assessment, in: Proc. Third International Conference on NETWORKING and Distributed Computing, 2012.

[14] S. Lu, J. Wang, H. Yang, H.-P. Zhang, Bayesian network model for fast disaster assessment in unconventional emergencies management, in: Proc. International Conference on Information Systems for Crisis Response and Management, 2012.

[15] B.-C. Li, J. Wang, C. Lin, Network public opinion early warning method based on intuitionistic fuzzy inference, Application Research of Computers 27(9)(2010) 3312-3315.

[16] L.-S. Shi, L. Chen, K. Li, A dynamic early warning method for network public opinion, Journal of Tianjin Normal University: Natural Science Edition 32(2)(2012) 59-65.

[17] H. Xing, X.-Y. Sun, M.-M. Wang, H.-Q. Zheng, Application of BP neural network and genetic algorithm in stress prediction of anchor bolt, in: Proc. 2015 7th International Conference on Modelling, Identification and Control (ICMIC), 2015.

[18] V. Gala, V. Deshpande, I. Ferwana, M. Milanova, Product sentiment trend prediction, in: Proc. International Conference on Social Computing and Social Media, 2018.

[19] R.A. Kamble, Short and long term stock trend prediction using decision tree, in: Proc. International Conference on Intelligent Computing and Control Systems, 2018.

[20] Y. Hu, Y.-M. Wang, Trend prediction method of Microblog public opinion based on fuzzy neural network, Information Science 35(12)(2017) 28-33.

[21] Z.-M. Zeng, C.-Y. Huang, Research on public opinion heat trend prediction model of emergent infectious diseases based on BP neural network, Journal of Modern Information 38(5)(2018) 37-44, 53.

[22] X.-L. Ye, K.-Y. Yang, Neural network optimization algorithm in public opinion trends prediction, Journal of Network New Media 5(1)(2016) 33-37, 51.

[23] X. Chen, M. Xia, J. Cheng, X. Tang, J. Zhang, Trend prediction of internet public opinion based on collaborative filtering, in: Proc. International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2016.

[24] Q.-L. Liu, J. Li, R.-B. Xiao, Trend prediction of public opinion propagation based on parameter inversion: an empirical study on Sina micro-blog, Journal of Computer Applications 32(5)(2017) 1419-1423.

[25] F.-L. Yin, B.-B. Zhang, Research on unequal time series clustering for hot topics, Journal of computer 29(4)(2018) 122-134.