The Application of Gray Model and Support Vector Machine in the Forecast of Online Public Opinion

http://dx.doi.org/10.3991/ijoe.v9iS2.2566

Liuwei Xu
Zhejiang Economic & Trade Polytechnic, Hangzhou, China

***Abstract*—The forecast of online public opinion is a kind of complex forecasting problem with information, small sample and uncertainty. In order to improve the accuracy for the forecast of online public opinion, a new forecasting method based on a gray model and a support vector machine is proposed. The method comprises the steps of clustering the text, extracting the hotspots, aggregating the data and implementing other pretreatments of the network data, then creating a model GM (1, 1) for the time series of online public opinion, correcting the forecasting results of the model GM (1, 1) with a support vector machine, and then testing through a simulation experiment. The experimental results show that compared with traditional forecasting methods, the application of gray model and support vector machine improves the accuracy for the forecast of online public opinion. Moreover, a new method for the forecast of online public opinion is presented to some extent.**

***Index Terms*—online public opinion, GM (1, 1), support vector machine, forecasting.**

I. INTRODUCTION

Online public opinion, which is also known as network public opinion, refers to the opinions or remarks with a certain influence and tendentiousness of netizen to the social public affairs, especially the hot social focuses through the Internet [1-2]. With the rapid development of Interne in China, network has become one of the main carriers for the reflection of social public opinion [3]. At present, the economic and social development of China is in a crucial stage, in which the various deeply rooted contradictions and problems arise day by day, so the hotspots of online public opinion are emerged one after another, which involve broad regions as well as extensive contents. In such a situation, the negative online public opinion will have great negative impact on the national security and social stability, if the online public opinion cannot be guided and supervised correctly [4]. Therefore, it has become a hotspot of research at present to forecast the trend of the development of online public opinion accurately.

In recent years, there are more and more studies focusing on the forecast of online public opinion, which can basically be divided into two categories: traditional forecasting method and modern forecasting method. According to the traditional forecasting method, the data of online public opinion is converted into time series, and the model is created by using the forecasting methods of autoregressive moving average, exponential smoothing and other time series. This method is simple and easy to be carried out. However, it assumes online public opinion is changed linearly, which is inconsistent with the actual changing characteristics, and therefore the results of forecast are not ideal. As for the modern forecasting method, the model is created on the basis of nonlinear theory. Compared with traditional forecasting method, the accuracy for the forecast of online public opinion is improved correspondingly, and the main forecasting models include Hidden Markov Model [5], G (Galam) [6], intuitionistic fuzzy reasoning [7], support vector machine [8-9], etc. Online public opinion is a kind of uncertain forecasting problem with information and small samples. In order to improve the accuracy of forecast further, some scholars have proposed some assembled forecasting models for the online public opinion based on the combination optimization theory and the advantages of each single model. For example, Zhang Jue put forward the online public opinion forecasting model based on ARIMA and BP neural network and achieved good forecasting results [10].

Gray forecasting theory [11] is proposed for the first time by domestic scholar, Deng Julong, in 1982, which studies “small sample” and “poor data information” uncertain system of “partial known information and partial unknown information. GM (1, 1) model, the important component of gray forecasting theory, is featured with less data required by model establishment. Support vector machine (SVM), referring to a modern machine learning algorithm specially for the small sample and uncertainty forecasting problems based on the statistical learning theory (SLT), is widely applied in the study of the field of nonlinear time series forecasting.

In the study, the grey model is attempted to be combined with the support vector machine model and applied in the forecast of online public opinion. Firstly, GM (1, 1) is used to establish the forecasting model of online public opinion. Secondly, the forecasting result of GM (1, 1) is modified by the support vector machine. At last the performance of the model is verified by simulation experiment.

II. PRETREATMENT FOR DATA OF ONLINE PUBLIC OPINION

For the information collection of online public opinion, we use the technology of network search robot, which can traverse the whole web space in the appointed range

to collect the various information sources of online public opinion thereon [12-13]. However, the messy and disordered data of online public opinion will be acquired, which should be converted into the related data by text clustering treatment.

A. Text Clustering

The hierarchical clustering algorithm [14-16] is used in the study to cluster the data of network pubic opinion, and the advantage and disadvantage of clustering is evaluated on the basis of the purity index. After the text clustering, the purity index for the clustering r is defined as follows:

$$P(S_r)=\frac{1}{n_r}\max\left(n_r^i\right) \tag{1}$$

In the formula, n_r is the number of documents in the r (th) clustering category, and n_r^i is the number of texts belonging to the predefined category i, but distributed into the r (th) clustering category by mistake.

So, all the purity indexes of the text clustering result are defined as follows:

$$purity=\sum_{r=1}^{k}\frac{n_r}{n}P(S_r) \tag{2}$$

B. Hotspot Acquisition

The hot topic of network means the information set regarding the network as the communications media, paid attention by a certain of crowd widely and continuously and capable of reflecting the situation of online pubic opinion [17-18]. The process of hotspot acquisition is as follows:

(1) The reporting frequency, continuous reporting time and network click rate of the topic are adopted as the characteristics of the hotspot topic, which are performed statistics.

(2) The values of media attention and public attention are calculated.

(3) The specific gravity balance factor and threshold value are set and the pubic attention is calculated

(4) If the public attention is more than the threshold value, it shows that the topic is the hotspot topic.

C. Data Aggregation

The collected online public opinion information with different vectors is organized by data aggregation and converted into the discrete-time series of hotspot topic by the data aggregation software.

III. FORECASTING MODEL OF ONLINE PUBLIC OPINION BASED ON GM (1, 1) AND SUPPORT VECTOR MACHINE

The gray model is capable of revealing the development trend of the data, but is not suitable for the forecast of time-invariance and nonlinear data, while the support vector machine is applied to describe the nonlinear and small sample data series. Thus, the forecasting model of online public opinion based on GM (1, 1) and support vector machine can be established by combining the advantages of the both.

A. GM (1, 1) Model

In recent years, the gray model GM (1, N) is widely applied and studied, wherein GM (1, 1), referring to the most common and simplest gray model and the model composed of differential equation only including single variable, is a special case of GM (1, M). Assume that the original data series is the model establishment series $X^{(0)}$ of GM (1, 1), that is to say:

$$X^{(0)}=\left(X^{(0)}(1),\ X^{(0)}(2),\Lambda,\ X^{(0)}(n)\right) \tag{3}$$

The original data series is accumulated by the accumulation and generation method, and 1-AGO series generated by accumulation for one time is as follows:

$$X^{(1)}=\left(X^{(1)}(1),\ X^{(1)}(2),\Lambda,\ X^{(1)}(n)\right) \tag{4}$$

In the formula, $X^{(1)}(1)=X^{(0)}(1)$, $X^{(1)}(k)=\sum_{m=1}^{k}X^{(0)}(m)$

Suppose $Z^{(1)}$ is the mean value series of $X^{(1)}$, $Z^{(1)}$ is obtained as follows:

$$Z^{(1)}(k)=0.5X^{(1)}(k)+0.5X^{(1)}(k-1) \tag{5}$$

So, the gray differential equation model of GM (1, 1) is as follows:

$$X^{(0)}(k)+aZ^{(1)}(k)=b \tag{6}$$

$k=2,3,\Lambda,n$ is put into the above formula and the formula is obtained as follows:

$$\begin{cases} X^{(0)}(2)+aZ^{(1)}(2)=b \\ X^{(0)}(3)+aZ^{(1)}(2)=b \\ \quad \text{M} \qquad \text{M} \quad \text{M} \\ X^{(0)}(4)+aZ^{(1)}(2)=b \end{cases} \tag{7}$$

The above equation can be converted into the matrix equation as follows:

$$Y_N=BP=\left[\left(X^{(0)}(2),X^{(0)}(3),\Lambda,X^{(0)}(n)\right)\right]^T \tag{8}$$

In the formula, B is the data matrix, Y_N is the data vector, P is the parameter vector, that is to say:

$$\begin{cases} B=\begin{bmatrix} -Z^{(1)}(2) & 1 \\ -Z^{(1)}(3) & 1 \\ \Lambda \quad \Lambda & \\ -Z^{(1)}(n) & 1 \end{bmatrix} \\ P=[a,b] \end{cases}$$

The solution is carried out by least square method to obtain the formula as follows:

$$P(a,b)^T=\left(B^TB\right)^{-1}B^TY_N \tag{9}$$

The obtained coefficient $P=[a,b]$ is put into the formula (6), and then the differential equation is solved to obtain the expression of gray GM (1, 1) intrinsic model as follows:

$$\hat{X}^{(0)}(k)=u^{k-2}\cdot v \tag{10}$$

In the formula, $u=\frac{1-0.5a}{1+0.5a}$, $v=\frac{b-a\cdot X^{(0)}(1)}{1+0.5a}$

Next, perform validation and suppose $\varepsilon(k)$ is the residual; that is to say:

$$\varepsilon(k)=\frac{X^{(0)}(k)-\hat{X}^{(0)}(k)}{X^{(0)}(k)}\times 100\% \qquad (11)$$

The residual is inversely proportional to the accuracy of model. For the general requirements, $\varepsilon(k)\le 20\%$, and the best condition is $\varepsilon(k)\le 10\%$.

B. Model of Support Vector Machine

The complexity of corresponding quadratic programming problem solved by the support vector machine is inversely proportional to the calculating speed. The least squares support vector machine (LSSVM), modifies the model of support vector machine and reduces the complexity of solution, so it has the advantages of less calculation resources as required as well as fast solution speed and convergence speed. Therefore, LSSVM is adopted as the forecasting mode in the study. For the time series of online public opinion, the regression function of LSSVM is as follows:

$$f(x)=w^T\varphi(x)+b \qquad (12)$$

In the formula, w is the weight vector and b is the bias constant.

According to the inductive principle of structure risk minimization, the model of least squares support vector machine for solving the regression problem is as follows:

$$\min_{w,b,\xi} L(w,b,\xi)=\min\frac{1}{2}\|w\|^2+\frac{1}{2}\gamma\sum_{i=1}^{n}\xi_i^2 \qquad (13)$$

The constraint condition is as follows:

$$y_i=w^T\varphi(x_i)+b+\xi_i \qquad i=1,\Lambda\ ,l \qquad (14)$$

In the formula, γ is the regularization parameter and ξ_i is the slack variable.

Lagrange multiplier is introduced to obtain the formula as follows:

$$L(w,b,\varsigma,a)=\frac{1}{2}w^Tw+\frac{1}{2}\gamma\sum_{i=1}^{l}\varsigma_i^2-\sum_{i=1}^{l}a_i(w^T\varphi(x_i)+b+\xi_i-y_i) \qquad (15)$$

In the formula, $a_i(i=1,\Lambda\ ,l)$ is Lagrange multiplier.

The following formula is obtained according to KKT (Karush-Kuhn-Tucker) condition in the optimization theory:

$$\frac{\partial L}{\partial w}=0,\frac{\partial L}{\partial b}=0,\frac{\partial L}{\partial \xi_i}=0,\frac{\partial L}{\partial a_i}=0 \qquad (16)$$

So, the last solution can be obtained as follows:

$$\begin{bmatrix}0 & E^T\\ E & R+I/\gamma\end{bmatrix}_{(l+1)\times(l+1)}\begin{bmatrix}b\\ a\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \qquad (17)$$

In the formula, $E=[1,\Lambda\ ,1]^T$, $y=[y_1,\Lambda\ ,y_l]^T$, $a=[a_1,\Lambda\ ,a_l]^T$, I is $l\times l$ step unit matrix, R is step matrix, and $R_{ij}=\varphi(x_i)^T\cdot\varphi(x_j)$. According to the Mercer condition, the kernel function is defined as follows:

$$K(x_i,x_j)=\varphi(x_i)^T\cdot\varphi(x_j)$$

The kernel function $K(x_i,x_j)$ is introduced to convert the formula (17) to obtain the forecasting model of LSSVM as follows:

$$f(x)=\sum_{i=1}^{l}a_iK(x_i,x_j)+b \qquad (18)$$

The radial basis function (RBF) is featured with good universality and better expression for processing the time series problem than that of other kernel functions, so in this paper, the radial basis function is used as the kernel function of LSSVM, and the expression is as follows:

$$K(x_i,x_j)=\exp\left(-\frac{|x_i-x_j|^2}{2\sigma^2}\right) \qquad (19)$$

In the formula, σ^2 is the kernel width of RBF.

IV. EXPERIMENTAL RESULT AND ANALYSIS

In order to verify the function of the gray model and support vector machine in the forecast of online public opinion, in the environment of Intel Core i5 3.2G CPU, 4GB RAM and hardware having Microsoft Windows Sever 2003 as the operating system, the implementation algorithm is realized by programming via MATLAB. A certain hot topic on the internet is forecasted and 30 data of the amount of the relevant posts is obtained, which is shown in Figure 1 for detail.

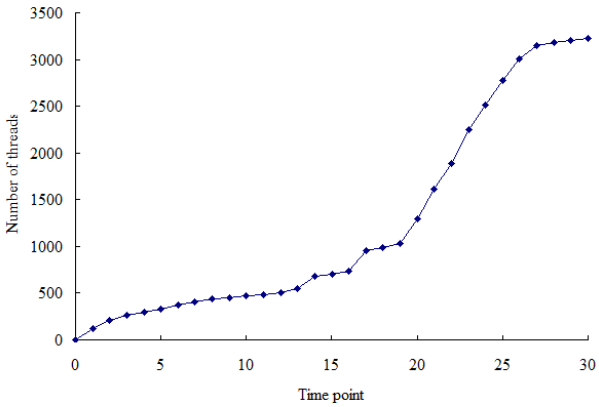

Figure 1. Time Series of Online Public Opinion

In order to quicken the training speed of the model to reflect the variation trend of the online public opinion better, the time series of online public opinion is pretreated, and the normalized data is shown below:

$$x_i'=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \qquad (20)$$

In the formula, x_i' is the data after normalization, $x_{\max}$ and $x_{\min}$ represent the maximum and minimum of the time series of online public opinion, respectively.

The data is divided into two parts. The former 22 data is used as the training sample set, and the later 8 data as the test sample set. And then the test sample set is forecasted by several models respectively. The obtained forecasting result is shown in Figure 2.

It is shown in the analysis for the value result of the test sample of online public opinion by the forecasting model in Figure 2 that the deviation between the forecasting

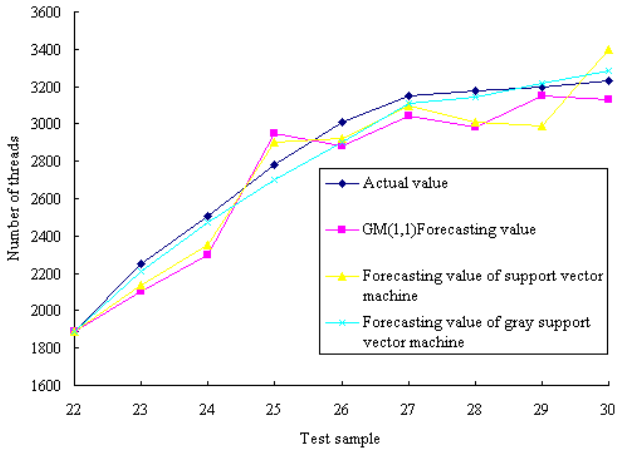

Figure 2. Forecasting Results for Test Sample by Several Models

results of the single gray model and support vector machine model and the actual value is large, and the errors are quite high, which indicates that a single model only can explain the fragment and part of variation rules of the complex online public opinion, but difficult to describe the laws of time invariance and nonlinear variation of online public opinion completely and accurately. However, the advantages of the forecasting models of the GM (1, 1) and support vector machine are combined with the advantages of the gray model and support vector machine to overcome the defect of the single model, which not only can explain the characteristics of set information and small sample of the online public opinion, and accurately forecast the rules of time invariance and uncertainty variation, but also capture the variation trend of the online public opinion, so as to improve the forecasting accuracy of it.

V. CONCLUSIONS

Online public opinion is a complex and time-varying system with larger burstiness and volatility. If it can be forecasted accurately, particularly, those hot public opinion which can arouse the attentions of most netizens, it will help the relevant departments to find out the potential risks timely, research and respond to the online public opinion actively, improve the ability to communicate with the public, and lead the online public opinion to the healthy development. In order to improve the accuracy for the forecast of online public opinion, a forecasting model of online public opinion based on the combination of a gray model and a support vector machine was proposed, aiming at the characteristics of online public opinion by taking advantages of the gray model and support vector machine. The experimental results show that the application of gray model and support vector machine can not only improve the accuracy for the forecast of online public opinion effectively and make up for the deficiency of a single forecasting model, but also provide a new idea for the study on the forecast of online public opinion.

REFERENCES

[1] Yang Yonghong, "Research of Net-Mediated Public Sentiment Based on Data Mining," M.S.thesis, College of Literature and Journalism, Chongqing University, Chongqing, China, 2010.

[2] Shasha Wang, "Phase Detection and Prediction Web Public Sentiment," in *Proceedings of the 2009 International Conference on Web Information Systems and Mining*, 2009, pp.116-117.

[3] Xu Xiaori, "Study on the Way to Solve the Paroxysmal Public Feelings on Internet," *Journal of North China Electric Power University(Social Sciences)*, No.1, 2007, pp.89-93.

[4] Wu Shaozhong, and Li Shuhua, "Research of Internet Public Opinion Early-warning Mechanism," *Journal of Chinese People's Public Security University (Science and Technology)*, No.3, 2008, pp.38-42.

[5] Marcelo Andrade Teixeira, and Gerson Zaverucha, "Fuzzy Hidden Markov Predietor in Electric Load Forecasting," in *Proeeedings of the 2004 IEEE International Joint Conference on Neural Networks*, 2004, pp.315-320.

[6] Serge Galam, "Minority Opinion Spreading in Random Geometry," *Eur.Phys.J.B*, Vol.25, No.4, 2002, pp.403-406. http://dx.doi.org/10.1140/epjb/e20020045

[7] Li Bicheng, Wang Jin, and Lin Chen, "Method of online public opinions pre-warning based on intuitionistic fuzzy reasoning," *Application Research of Computers*, Vol.27, No.9, 2010, pp.3312-3315.

[8] Vapnik V, *The Nature of Statistical Learning Theory*, New York: Springer, 1995.

[9] Quanlong Guan, Saizhi Ye, Guoxiang Yao, Huanming Zhang, Linfeng Wei, Gazi Song, and Kejing He, "Research and Design of Internet Public Opinion Analysis System," *2009 IITA International Conference on Services Science, Management and Engineering*, 2009, pp.173-177. http://dx.doi.org/10.1109/SSME.2009.62

[10] Zhang Jue, "Research on Forecasting Models and Platform of Online Public Opinion," M.S. thesis, Beijing Jiaotong University, Beijing, China, 2009.

[11] Chaang-Yung Kung, Huei-Shr Chen, Chih-Cheng Hang, and Kun-Li Wen, "Using GM(l, l) Method to Foreeast the Development of Cell Phone Market in Taiwan," in *Proeeedings of the 2008 IEEE International Conference on Fuzzy Systems*, 2008, pp. 539-543.

[12] Pang-Ning Tan, and Vipin Kumar, "Discovery of Web Robot Sessions Based on their Navigational Patterns," *Data Mining and Knowledge Discovery*, Vol.6, No.1, 2002, pp.9-35. http://dx.doi.org/10.1023/A:1013228602957

[13] Derek Doran, and Swapna S. Gokhale, "Web robot detection techniques: overview and limitations," *Data Mining and Knowledge Discovery*, Vol.22, No.1-2, 2011, pp.183-210. http://dx.doi.org/10.1007/s10618-010-0180-z

[14] Li B, Chen Y, and Bai X, "Experimental study on representing units in Chinese text categorization," in *Proceedings of CICLing*, 2003, Springer, pp.602-614.

[15] Lwis D D, and Hayes PJ, "Guest editorial-special issue on text categorization," *ACM Transactions on Information Systems*, Vol.12, No.3, 1994, pp.231.

[16] Kumar P, Krishna PR, Bapi RS, and De SK, "Rough clustering of sequential data," *Data & Knowledge Engineering*, Vol.3, No.2, 2007, pp.183-199. http://dx.doi.org/10.1016/j.datak.2007.01.003

[17] Li Juan, Zhou Xueguang, and Chen Bin, "Research on Analysis and Monitoring of Internet Public Opinion," in *Proceedings of the 2012 International Conference of Modern Computer Science and Applications*, 2012, Vol.191, pp.449-453. http://dx.doi.org/10.1007/978-3-642-33030-8_72

[18] Hong Liu, and Xiaojun Li, "Internet Public Opinion Hotspot Detection Research Based on K-means Algorithm," in *First International Conference ICSI*, 2010, Vol.6146, pp.594-602.

AUTHORS

Liuwei Xu is a lecturer in Zhejiang Economic & Trade Polytechnic, Hangzhou, China (wuwei6868@ gmail.com).

This article is an extended and modified version of a paper presented at the International Conference on Mechanical Engineering, Automation and Material Science (MEAMS2012), held 22-23 December 2012, Wuhan, China. Received 14 February 2013. Published as resubmitted by the authors 25 March 2013.