Study on the pre-warning system of public opinion based on PCA and PSO-SVM

Xiaohong Hao[1, a], Kaicheng Gu [1, b], Boyu Meng[1, c]

[1]School of Computer & Communication, Lanzhou University of Technology, Lanzhou 730050, China

[a]316475958@qq.com, [b]gkc1314@qq.com, [c]boyu8816@163.com

Keywords: Index System, Principal Component Analysis, Support Vector Machine, Particle Swarm Optimization.

Abstract. In "Internet+" era, how to real-time monitoring and build effective pre-warning systems of the network public opinion crisis has become a required courses for government departments and enterprises. This paper sufficiently considers the development， changes in laws and characteristics of the network public o-pinion crisis, and establishes a pre-warning index system with 7 indexes of network public opinion. Because of the final index redundancy and the lack of data, building the public opinion of Pre-Warning model based on PCA and SVM, and using PSO to optimize the parameters of SVM. Experiment shows the pre-warning index system of network public opinion and the model of PCA-PSO-SVM is effective and feasible.

Introduction

Along with the rapid development of Internet, the Internet is becoming more and more open, hot topic as a form of network public opinion, has the characteristics of short-term, sudden and real-time, and if lack of the necessary preventive measures, they will quickly erupt and generate huge influence[1]. Therefore, it is necessary to construct the pre-warning system of the network public opinion.

Main work

About the index system of public opinion, researchers' professional background and the viewing angle are differences, so formed many different index systems. Although these index systems considering the influence of subject factors as far as possible, but still exist some shortcomings [2]:

Therefore, this paper considers the disadvantages of the current index system of public opinion, according to the characteristics of timeliness, fast reaction, from the acquisition mode of data, construct a simple and feasible index system of public opinion. This paper establishes a pre-warning index system with 7 indexes of public opinion.

(X1)Attention: this parameter reflects users' attention of different topics in different time, directly and objectively reflect the interest in the topic and the demand of internet users. $X_1 = \frac{t}{T}$, t is daily search index, T is topical highest search index. Both of them are from baidu index.

(X2)Participation: the parameter reflects users' participation for different topics in different time. $X_2 = \frac{d}{D}$, d is daily weibo, D is weibo highest amount. Both of them are from sina weibo

(X3)Intuitionistic degree: The study found a topic with pictures, video or sound that spread more quickly. $X_3 = \frac{d_v}{d}$, d_v is daily weibo with video and images.

(X4)Integrate degree: the parameter reflects the integrate degree of the topic. $X_4 = \frac{d_l}{d_o}$, d_l is daily original weibo with links, d_o is daily original weibo.

(X5)Influence degree: because the verified account has more fans, they published content can

have a greater impact. $X_5 = \frac{d_{ov}}{d_o}$, d_{ov} is daily original weibo of verified account.

(X6)Change of attention: the parameter reflects the change of Internet users to the awareness of topic. $X_6 = X_1 - X'_1$, X_1' is the Attention of the day before.

(X7)Change of participation: the parameter reflects the change of Internet users to the participation of topic. $X_7 = X_2 - X'_2$, X_2' is the Participation of the day before.

However, there are many factors of public opinion. In the index system, there are seven indexes, in the process of perfecting the index system will further increase the final index number, and this will no doubt greatly increase the amount of calculation, results in the decrease of the accuracy and efficiency of model. So principal component analysis (PCA) is introduced for reduction of attributes, reduce the dimension of the original data to simplify data structure, can make many original data has certain correlation into a set of new index has nothing to do with each other.

Introduce SVM into the field of public opinion, by comparing the experimental research of the parameters of SVM, there are practical significance to the study of public opinion. For the type of kernel function, including polynomial kernel and radial basis function kernel(RBF) and Sigmoid function, the practice shows that RBF can make SVM to achieve the best result, so this paper use RBF as the kernel function of SVM [3]. For ε and C, Chen and Zheng [4] used different generalization estimates as the fitness function of genetic algorithm (GA), they proposed two kinds of the SVM parameter selection based on GA. But the GA is more complex, Particle Swarm Optimization (PSO) is different with GA with "survival of the fittest, superior bad discard", the concept of it is simpler, more efficient. To solve the optimization problem, the solution of the problem corresponding to the location of the search space of a bird, called the birds "particles". Each particle has its own position and velocity, there is a fitness value decided by be optimized function. All particles memorize and following current best, search PSO in the solution space is initialized to a random particles, and then to find the optimal solution by iteration.

Particles by track two "extremes" for renew itself. One is the particle find the optimal solution by itself, the solution is called individual extreme P_{Best}, the other is the optimal solution found by entire population G_{Best}.when find the two best value, particles according to the following formulas to update their speed and location:

$$V_{id}^{k+1} = \omega V_{id}^{k} + c_1 r_1 (P_{id}^{k} - X_{id}^{k}) + c_2 r_2 (P_{gd}^{k} - X_{id}^{k}) \quad (1)$$

$$X_{id}^{k+1} = X_{id}^{k} + \beta V_{id}^{k+1} \quad (2)$$

ω is non-negative number, called inertia weight; $d = 1, 2, \cdots\cdots, \mathrm{D}$; $i = 1, 2, \cdots\cdots, \mathrm{n}$; k is the current number of iterations; V_{id} is the speed of the particle; β called constraint factor, Controlling the speed of weight, usually take 1; c_1 and c_2 are the constant non-negative, called learning factors, usually take 2; r_1 and r_2 are the random number between distributed in (0, 1).

Different types of hot topic have a fixed lifetime, Researchers to study the life cycle think that, Gestation period: the positive and negative energy accumulated, triggering public opinion crisis more likely; Outbreak period: the attention of Internet users is highest, the accumulation of positive and negative energy released that most harmfulness to society; Remission period: decline concerns of Internet users, but positive and negative energy is still at a high level; Fade period: users' attention is low and energy release out. This paper take the Gestation period (A), Outbreak period (B), Remission period (C) and Fade period (D), corresponding to the four different warning levels, namely the heavier, especially serious, serious, generally, they are used to test the feasibility of the model. Pre-warning model of public opinion is designed as shown in Figure 1.

The selection of SVM parameters based on PSO described as follows:

Step 1 Read sample data, randomly generates a set of $\{C, \varepsilon\}$ as the particle's initial position;

Step 2 Average divided the whole sample into K subsets $S_1, S_2, S_3, \ldots, S_k$ which each other does

not contain; Step 3 According to the current { C , ε } training SVM, computing k-fold cross-validation error;

Step 3.1 Initialize i=1; Step 3.2 S_i are reserved for testing set, and the rest merged as the training set, training the SVM; Step 3.3 Calculate classification error of the i subset, making i=i+1, repeat steps 3.2 until i=K+1; Step 3.4 Calculate the K average classification error and get the k-fold cross-validation error;

Step 4 K-fold cross-validation error as fitness, and remember the best fit for the individual and group values P_{Best} G_{Best} , according to the (1)and (2) to search for better { C , ε }; Step 5 Repeat Step 2 until the number of iterations; Step 6 End.

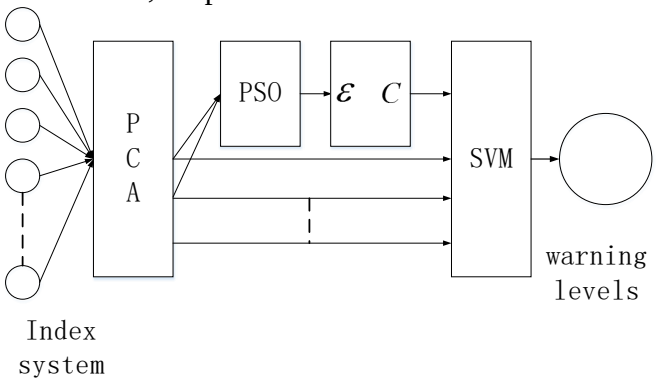

Fig. 1 model based on PCA and PSO-SVM

Experiments and analysis

According to the index system, collecting case (1) "Maoming PX event", case (2) "Zhejiang Cangnan Chengguan hit event", case (3) "Zhaoyuan murder", case (4) "Fu Hi group", case (5) "Malaysia airlines aircraft missing", from Sina weibo and Baidu index. Data from different life cycle total 37 samples. After pretreatment, using SPPS Version22 data analysis software to process the 37 samples, PCA is used to subtract the number of input factors. Analysis shows the first 5 main components concentration about 95.339% of the original seven variable, in table 1.

Randomly selected 11 as testing samples, the rest of the samples as training samples. Test samples and training samples contain each stage of development, using the GA-SVM, PSO-SVM, and 3-fold cross-validation. After training samples to determine C , ε the pre-warning results by SVM are shown in table 2. Respectively GA-SVM, PSO-SVM, PCA-GA-SVM, PCA-PSO-SVM four models are compared. To the pre-warning system of public opinion, PSO and GA are able to further improve the accuracy of the results. In addition, because of the PCA in advance reduce the number of index system for GA-SVM and PSO-SVM, not only improves the accuracy of the models, and time shortened respectively by 46.68%, 14.78%. Confirmed under the prerequisite of ensuring the accuracy of warning, based on PCA and PSO-SVM joint methods for pre-warning system of public opinion is the most effective and feasible.

Table.1 principal component analysis

Component	Total	%of Variance	Cumulative%
1	0.245	40.150	40.150
2	0.123	20.182	60.332
3	0.095	15.560	75.893
4	0.074	12.071	87.964
5	0.045	7.375	95.339
6	0.016	2.613	97.953
7	0.012	2.047	100.00

Table.2 Experimental results

Algorithms	C	ε	Accuracy of training	Accuracy of testing	Time
GA-SVM	6.57031	0.308594	96.1538%	63.6364%	12.007483s
PSO-SVM	3.36359	0.594604	92.3077%	72.7273%	10.078786s
PCA-GA-SVM	3.7968	0.13294	88.4615%	72.7273%	6.402475s
PCA-PSO-SVM	1.5	1.7	88.4615%	90.9091%	8.589622s

Conclusion

This paper builds the pre-warning model of public opinion based on the index system and PCA-SVM, using PSO to optimize the parameters of SVM, further improve the reliability of the model. Through the experiment shows that the index system of public opinion and pre-warning model are effective and feasible, which can help government departments and enterprises to analyze the public opinion and provide the basis for future efforts to achieve real-time pre-warning.

Reference

[1] Sun Ling-fang，ZHOU Jiabo，LIN Weijian. On Network Public Opinion Crisis Early Warning Based on the BP Neural Network and Genetic Algorithm[J].Journal of intelligence，2014，33（11）：18-24.

[2] Tan Guo-xin. Public emergency network public opinion research on monitoring index system[J]. Journal of Huazhong Normal University，2010，49（3）：66-70.

[3] Gil－Garcia R，Pons－Porrata A．Dynamic Hierarchical Algorithms for Document clustering[J]．Pattern，Recognition Letters，2010，(31):469－477．

[4] Chen Peng-wei, WANG Jung-ying. Model selection of SVMs using GA approach[C] Proc of 2004 IEEE Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE Press, 2004:2035-2040