Prediction Model based on Internet News Buzzword Data

Xuan Lei

School of Control and computer Engineering, North China Electric Power University, Baoding
071000, China
2582171339@qq.com

Abstract. The partial classification algorithm is mainly used to predict the popularity of network news and to explore the best model to predict the popularity of network news, so as to help network news service providers predict the popularity of news before publication. The popularity of network news is predicted according to the data analysis process: first, UCI data sets are pre-processed; secondly, feature selection is conducted for the data sets by using recursive feature elimination algorithm; then modelling and analysis is carried out, and finally through the confusion matrix, risk map and ROC (Receiver Operating Characteristic) chart performance evaluation, the performance of the model is compared and analyzed. Through comparison, it is found that random forest is the best prediction model.

Keywords: network news; feature selection; classification algorithm; model evaluation.

1. Introduction

The content production and exchange platform based on user relationship on the Internet is called social media, such as Zhizhi, Sina Weibo, Weixin, Twitter and so on. Social media allows users to publish text, pictures, audio and video, and other forms of content, users can share the content published by others and spread through the network composed of user relationships. The growing user groups and social media give users this autonomy. As a result, users publish more and more content, which has brought explosive growth in the amount of information.

Nowadays, network news service has become the mainstream network information carrier, and is also an indispensable channel for people to obtain news information in their daily life. Traditional news acquisition methods such as newspapers, radio and television news have gradually been replaced by network news. Network news has the characteristics of timeliness and convenience, and has won the favour of the masses of the people. It has become a habit for people to browse the news on the Internet every day. People can use electronic devices to learn the news in real time without spending money on newspapers. Thousands of online news are constantly available to network users, and news service providers must compete for readers, especially the time and energy of readers. By understanding and looking for readers' interests, news service providers can stand out and gain more benefits. Readers clicking on or not clicking on news articles are influenced by many factors, including the location of the article on the page, the time it was published, the topic of the article, the content of the article and so on. People click on the Internet news basically through the limited information, and the news successfully gain readers' attention can become interesting and popular news.

2. Literature Review

In recent years, the hotspots of social media research are as follows:

Statistical analysis of complex networks. Using user data and user relationship data on social media, a complex network is abstracted as the research object to study how to use a known probability distribution to better fit the degree distribution of nodes; to study the evolution of nodes and edges in complex networks; to study the measurement indicator of the importance of edges and nodes; to study how to extract a sub-network as the basic framework of the network and keep some statistical characteristics unchanged; and to study the robustness of network [1-2].

Recommendation system. Users and messages on social media are regarded as user node set and message node set, respectively, and there is no connection between user node set and any two nodes in message node set. Users' choice of messages is regarded as the side connecting the two node sets, forming a user-message bipartite graph network. Based on the network, many recommendation algorithms are proposed, such as user-based collaborative filtering algorithm, commodity-based collaborative filtering algorithm, substance diffusion algorithm and heat conduction algorithm [3-4].

The study of popularity prediction can be traced back to the user's attention to page access. Statistical research shows that the distribution of page visits is uneven, a few pages get a large number of visits, and most pages only get a small number of visits, that is, user access to the page is biased. Zipf law can be used to describe this phenomenon [5].

These studies show that content on social media has a short lifecycle and can only attract users' attention in the short term, and it is difficult to predict its future popularity. Later, Albert et al. found that there was a positive linear relationship between the future popularity of content and its early popularity on through the study of data on YouTube and Digg datasets [6]. Based on this, a constant scaling model was proposed, that is, a constant coefficient was amplified according to the early popularity of content after publication to get the later popularity prediction. This method has achieved good results in the prediction of content popularity of the two short life cycles. With the mining of more relevant information and application of feature engineering, classification-based popularity prediction method is proposed to predict which numerical range the future popularity will be [7]. Lammers et al. extracted content features and context features (number of followers, number of concerns, length of account creation, number of favourite tweets, and total number of historical forwards) from Twitter datasets, used principal component analysis to analyze the relationship between forwarding numbers and content features and context features, and used generalized linear model to analyze to what extent the content features and context features will affect the forwarding behaviour [8]. Vergouwe et al. put forward the SEHP model and believed that the growth rate of popularity was the cumulative effect of initial release and subsequent forwarding [9].

3. Case Analysis

3.1 Data Pre-processing and Feature Selection

For data analysis, the original data obtained may not be directly used in general, but must be pre-processed, because the original data set may be mixed with noise, missing values, redundant repetition and so on. Data cleaning is completed according to data analysis requirements, such as judging missing values, processing noise, and data transformation [10].

The recursive feature elimination algorithm established based on model-based feature selection method is used here. The main idea of recursive feature elimination is to iteratively construct a model, then select the best (or worst) features according to the certain evaluation criteria, put aside the selected features that meet the requirements, and repeat the process on the remaining features until all the features are traversed. In this process, the order of feature elimination is the sorting of features. Therefore, it is also a greedy algorithm for finding the best characteristic subset. In practice, recursive feature elimination (function) of care package is used to test independent variables of different numbers. Then, the control parameters are defined, the independent variables with different models are sorted, and the prediction accuracy of different numbers of independent variables is observed. According to the comparison, it is found that the accuracy of the model is the highest when selecting 20 variables. Therefore, 20 variables are selected as the final result of feature selection.

3.2 Adaptive Enhancement Method Analysis

The adabag package of R software is adopted to implement the algorithm. Figure 1 below shows the relationship between the number of iterations and model errors, in which the vertical axis is the value of the specific error rate and the horizontal axis is the number of decision trees. Through observation, it can be found that with the increase of the number of iterations, the training error rate decreases gradually. When the number of iterations in the model is less than 100, the training error of the model will fluctuate greatly. When the number of iterations is greater than 100, the training error of the model will tend to be stable, but there are still some changes, and when the number of iterations is close to 500, the model error no longer changes.

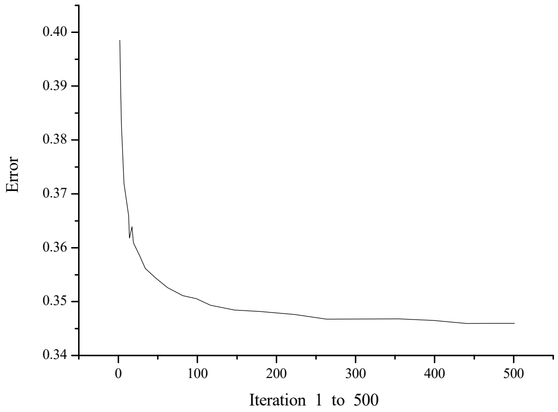

Figure. 1 Relationship between iteration number and model error

The prediction results based on the OOB sample set are shown in Table 1 of the confusion matrix. The total training error of the model is 34.4%. When the number of iterations is 478, the out-of-pocket error is 34.5%. The difference between the prediction result of the final model and the actual result of the training set is shown in the following table. It can be seen that the final model correctly predicts 9223 training set samples in category 0 (not popular) and wrongly predicts 4799 samples as category 1 (popular) incorrectly. The prediction misjudgement rate under this kind of samples is 34.23%; the final model correctly predicts 9005 training set samples in category 1 (popular), and wrongly predicts 4727 samples as category 0 (not popular), and the prediction misjudgement rate is 34.43% under such samples.

Table 1. AdaBoost model confusion matrix result table

		True category	
		0 (not popular)	1 (popular)
Forecast category	0 (not popular)	9223	4727
	1 (popular)	4799	9005

3.3 Random Forest Algorithm Analysis

The random Forest package of R software is used to analyze the characteristics of the popularity degree in the data set, and establish a suitable random forest model. The model constructed is also analysed to see how the prediction ability of the model is.

Through observation, it is found that the minimum error of the model occurs when the number of decision trees is 500. Therefore, the classification of the model constructed is discriminant model. The random forest model contains 500 decision trees. The number of predictive variables considered at each splitting point is about equal to the square root of the total number of predictive variables, that is, 5. The prediction results based on OOB sample set are shown in the confusion matrix table, and the total prediction error of the model is 35.49%. The difference between the prediction result of the final model and the actual result of the training set is shown in Table 2. As can be seen from Table 2, the final model correctly predicts 9151 training set samples as category 0 (not popular), wrongly predicts 4871 samples as category 1 (popular), and the prediction misjudgement rate under this type of sample is 34.74%; and the final model correctly predicts 8754 training set samples as category 1 (popular), of which 4978 sample are erroneously predicted to be category 0 (not popular), and the prediction misjudgement rate is 36.25% under such samples.

Table 2. Random forest model confusion matrix result table

		True category	
		0 (not popular)	1 (popular)
Forecast category	0 (not popular)	9151	4978
	1 (popular)	4871	8754

Figure 2 below shows the ROC (Receiver Operating Characteristic) image calculated and plotted from the out-of-pocket data of the training set in the random forest model. The hit ratio of the vertical axis is the ratio of the correct number of positive cases to all positive cases. The false alarm ratio of the horizontal axis is the ratio of the number of negative cases to all negative cases. The AUC (Area under Concentration-time Curve) value of the training set is 0.72.

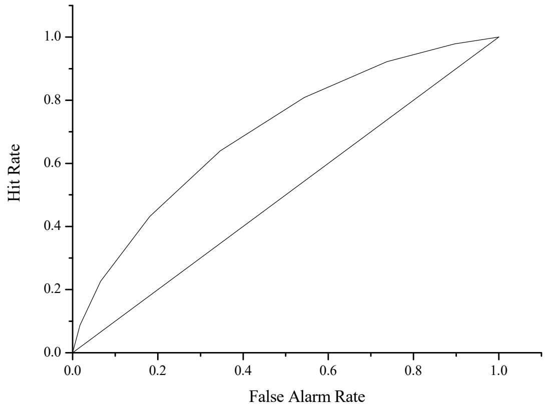

Figure. 2 ROC diagrams of random forests

3.4 Support Vector Machine Algorithm Analysis

The kernlab package of R software is used for modelling and analysis of support vector machines. The kernel functions in the modelling process of support vector machines include linear kernel function, polynomial kernel function and radial basis function (also known as Gaussian function). Different kernel functions are used for modelling analysis, and the results can be obtained as shown in Table 3.

Table 3. Results of support vector machines with different kernel functions

Kernel function type	The number of support vectors	Training error	Test error
Linear kernel function	21793	0.358992	0.359281
Radial basis function	20944	0.342668	0.348353
Polynomial kernel function (d=2)	20906	0.360938	0.354446
Polynomial kernel function (d=5)	17245	0.340758	0.437795
Polynomial kernel function (d=7)	12186	0.492047	0.486047
Polynomial kernel function (d=9)	12033	0.494740	0.489914
Polynomial kernel function (d=10)	11898	0.440216	0.459020

Through Table 3, it is seen that the training error and test error of linear kernel support vector machine are higher than that of radial basis function. Continuing to use more complex kernel functions to train support vector machine, it can be found that the training error of the model increases continuously, and the training error reaches the maximum when the degree of freedom of the polynomial kernel function is 9 and then begins to decline. Moreover, when the degree of freedom of the polynomial kernel function is 10, the test error is much larger than the training error, which can be said to produce over-fitting. In general, the performance of the support vector machine model is the best when the number of radial basis function cores is taken. The number of training sets is continuously adjusted and the training error and test error of different training sets are calculated. Then, it is found that the training error increases with the increase of number of training sets, but the training error fluctuates within a certain range after a certain number, and the test error does not change much after a certain number. Moreover, the value of different numbers of training errors and test errors are not necessarily.

4. Model Evaluation

First, comparison of confusion matrices between different models. The confusion matrix between the models is shown in Table 4.

Table 4. Confusion matrix table

Random forest model		
	Predicted	
True	0	1
0	2008	1032
1	1039	1873
Support vector machine model		
	Predicted	
True	0	1
0	1989	1051
1	1023	1889
Adaptive enhancement model		
	Predicted	
True	0	1
0	1985	1055
1	1007	1905

It can be found from Table 4 that the prediction error of the random forest model is 34.80%, that is, 17.34% of the samples whose real results are zero are mistakenly predicted as the category 1, and 17.46% of the samples whose real results are 1 are mistakenly predicted as the category 0; the prediction error of the support vector machine model is 34.86%, that is, 17.66% of the samples whose real results are zero are mistakenly predicted as category 1, and 17.19% of the samples whose true results are 1 are mistakenly predicted as category 0; the prediction error of adaptive enhancement model is 34.64%, that is, 17.73% of the samples with true results of 0 are mistakenly predicted as category 1, and 16.92% of the samples with true results of 1 mistakenly predicted as category 0.

Simply analyzing the confusion matrix of the prediction model, it can be found that the error rates of the support vector machine model and the random model are lower than those of the adaptive enhancement model. Nevertheless, in general, the prediction error rates of the three models are not very different, all below 35%. That is to say, from the perspective of error rate alone, it is impossible to judge which model is the optimal.

Second, the analysis of risk maps between different models. When understanding the risk maps, the specific scenario is to analyze the process. Through the analysis, it is found that for a 50% sample ratio, the model performance will be reduced to less than 70% of the original model. Through observation, it can be found that 80% of the risk maps drawn by the adaptive selection model and the random forest model are below the Recall line. On the other hand, the area under the Recall line of the support vector machine model is 79%. Therefore, the data analysis and prediction are: the use of random forest model and adaptive selection model is the most economical and effective.

Third, ROC diagrams and related charts between different models.

The variation of errors among random forest model, adaptive selection model and support vector machine model can be seen concretely from figs. 2, 3 and 4 above. The relationship between correct positive judgment rate and false positive judgment rate is drawn in the ROC image. The maximum AUC value of random forest is 0.72. That is to say, the random forest model has the best prediction effect.

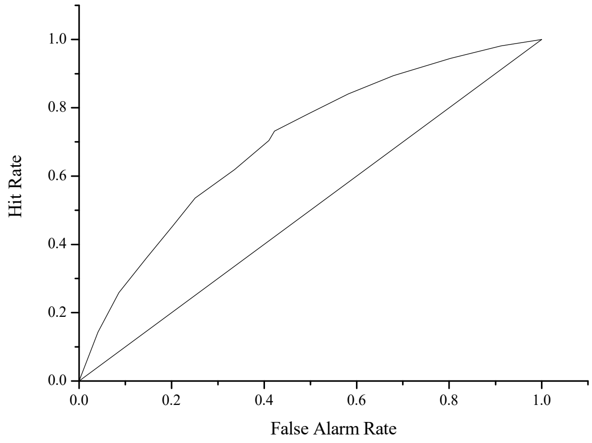

Figure. 3 ROC diagram of adaptive selection model

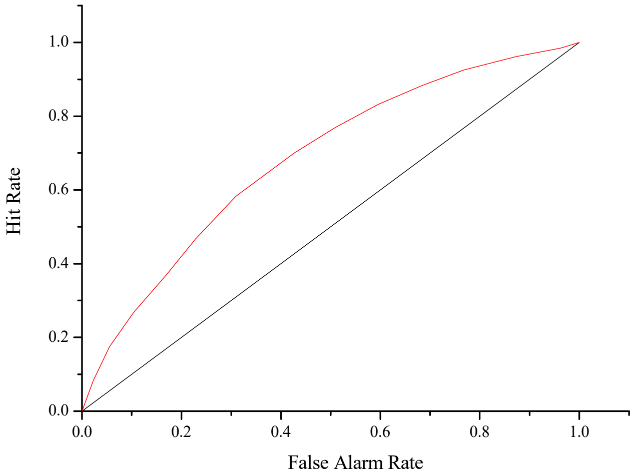

Figure. 4 ROC diagram of support vector machine model

5. Conclusion

Three models, adaptive selection, random forest and support vector machine, are used to analyze the data of the prediction of the popularity degree of network news and each model has many parameters. In the process of modelling, it is found that the optimal number of decision trees in the random forest model is 500. The optimal iteration number of the adaptive selection model is 500. When the radial basis kernel function is taken, the performance of the support vector machine model is the best. In the process of model evaluation, the confusion matrix table is firstly calculated, but by comparing the results of the confusion matrix, it is found that the test error rates of the three models are not very different. It is unable to directly select the optimal model, and then from the perspective of efficiency, the risk maps of the three models are compared and found that choosing random forest model and adaptive selection model is the most economical and effective. However, it is still unable to choose the best model, so the ROC diagram of the three models is observed and the AUC value is compared. It can be found that the highest AUC value of random forest is 0.72, that is, the random forest model has the best prediction effect, so the model chosen at last is random forest model.

References

[1]. Jaiswal, R. K., & Jaidhar, C. D. (2016). Location prediction algorithm for a nonlinear vehicular movement in vanet using extended kalman filter. Wireless Networks, 23(7), 1-16.

[2]. Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., & Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector

machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides, 13(2), 361-378.

[3]. Schneider, J., Hapfelmeier, A., Sieglinde Thöres, Obermeier, A., Schulz, C., & Nennstiel, S., et al. (2016). Mortality risk for acute cholangitis (mac): a risk prediction model for in-hospital mortality in patients with acute cholangitis. Bmc Gastroenterology, 16(1), 1-8.

[4]. Yu, B., Zhu, Y., Wang, T., & Zhu, Y. (2016). A 10-min rainfall prediction model for debris flows triggered by a runoff induced mechanism. Environmental Earth Sciences, 75(3), 216.

[5]. Heus, P., Damen, J. A. A. G., Pajouheshnia, R., Scholten, R. J. P. M., Reitsma, J. B., & Collins, G. S., et al. (2018). Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the tripod statement. Bmc Medicine, 16(1).

[6]. Albert, R. A. R., Biederman, J. Diepstraten, J. Anouk, D. A. D. Kruip, M. & Bemt, P. V. D. et al. (2018). Development of a clinical prediction model for an international normalised ratio >= 4.5 in hospitalised patients using vitamin k antagonists. British Journal of Haematology, 181(1).

[7]. Peltan, I. D. Rowhani-Rahbar, A. Vande Vusse, L. K. Caldwell, E. Rea, T. D. &Maier, R. V.et al. (2016). Development and validation of a prehospital prediction model for acute traumatic coagulopathy. Critical Care, 20(1), 371.

[8]. Lammers, R. J. M., Hendriks, J. C. M., Witjes, W. P. J., Palou, J., & Witjes, J. A. (2016). Prediction model for recurrence probabilities after intravesical chemotherapy in patients with intermediate-risk non-muscle-invasive bladder cancer, including external validation. World Journal of Urology, 34(2), 173-180.

[9]. Vergouwe, Y. Nieboer, D. & Oostenbrink, R. (2016). A closed testing procedure to select an appropriate method for updating prediction models. Statistics in Medicine, 36(28), 4529-4539.

[10]. Hristov, A. N., Kebreab, E., Niu, M., Oh, J., Bannink, A., & Bayat, A. R., et al. (2018). Symposium review: uncertainties in enteric methane inventories, measurement techniques, and prediction models. Journal of Dairy Science, 101(7), 6655.