EasyChair Preprint

№ 739

Influence Evaluation Model of Microblog User Based on Gaussian Bayesian Derivative Classifier

Zhe Zheng, Chunliang Zhou and Weipeng Zhang

January 19, 2019

Influence Evaluation Model of Microblog User Based on Gaussian Bayesian Derivative Classifier

Zhe Zheng[1] and Chunliang Zhou[2*] and Weipeng Zhang[2]
[1]Ningbo city college of vocational technology, Ningbo,China
[2]Ningbo University of Finance & Economics, Ningbo,China
zhengzhe@nbcc.cn,chunliangzhou@nbdhyu.edu.cn,zwphb@163.com

Abstract

In order to depict the influence of Weibo user, an evaluation model is proposed with Gaussian Bayesian derivatives. At first, the influence indexes of Weibo user is presented in this model with activity degree, relation degree and coverage degree. Combining the relationship characteristics between users and behavioral characteristics of user, the solved method for this model is given by Gaussian Bayesian derivatives. At last, a simulation is conducted to study the influence factor with experiment data of Sina Weibo users. The results show that, compared to other algorithm, this method has good adaptability.

Keywords: Weibo User; Influence; Gaussian Bayesian Derivatives

1 Introduction

As an important information promotion channel, microblog has a huge impact on social life[1-5].However, users are the foundation of microblog relationship, so the developed microblog user network has become a common social network. The greater the influence of microblog user is, the greater the role in information dissemination[6-9]. On the one hand, microblog provides convenience for people's daily life; on the other hand, it also brings adverse impact to the society. For example, rumors spread through microblog, and their influence scope and spreading speed are huge.

Therefore, scholars at home and abroad have conducted a lot of researches on the influence of microblog users. Mao Jiaxin et al[10] proposed a method to analyze and measure the social influence of users by predicting their ability to spread information based on the dissemination situation of social influence in microblog and in combination with the social network structure and user behavior factors, and obtained better influence estimation results. Wang Yaqi et al[11] proposed an evolution model of microblog user relationship network based on the internal characteristics of microblog user relationship network and the mean field theory to deeply analyze the dissemination dynamic behaviors of rumors on such evolution model and the topological statistical characteristics of such evolution model, and the results show that the microblog user relationship network has scale-free characteristics, and the degree distribution index is not only related to the probability of reverse connection, but also related to the node attraction degree distribution. In order to effectively depict the microblog user relationship, Xu Zhiming et al[12] gave the calculation method of user similarity based on various user attribute information by taking the microblog social network as weighted undirected graph, and the experimental results show that the user similarity based on social information has good adaptability in user relationship analysis and group mining. On the basis of K2 algorithm, Liu Haoran et al[13] proposed a new Bayesian structure learning algorithm which establishes the maximum spanning tree and gets the maximum number of parent nodes by calculating mutual information, and simultaneously searches the maximum spanning tree using ant colony algorithm to obtain the node order, and finally gets the optimal Bayesian network structure as per K2 algorithm. The experimental results show that this method solves the problem that K2 algorithm relies on prior knowledge and simplifies the search mechanism. Zhu Guofeng et al[14] proposed a method to calculate the influence of micro-blog users in combination with the intersectionality characteristics of microblog field. This method identifies the field of microblog based on the similarity between the user's tweets and the field itself, and calculates the influence of users in each field according to the user attributes respectively, and thus determines the influence of microblog users. Guo Hao et al[15] proposed a method to quantify the user's influence based on the scope of information dissemination on account of the quantification problem of user's influence. Meanwhile, by comparing real data sets and experimental results, the results show that compared with other measurement methods, this method is suitable for environments where data sets and time periods need to be limited, and the computational complexity is lower. Zhang Shaowu et al[16] established a method to measure the microblog influence, behavior influence and activeness influence based on the traditional influence measurement index and in combination with the user activeness, microblog value and message dissemination influence diffusion, and proposed the influence measurement model of microblog users based on these three new measurement methods.

However, the current research has not fully considered the intersectionality characteristics of microblog field, and many researches are not based on the microblog user relationship network, which greatly reduces the practicality and reference value of the results. As a result, by combining Gaussian Bayesian Derivative Classifier[17-20], this paper proposes a model to evaluate the influence of microblog

users, which first proposes the influence depiction index of microblog users, and then gives the method to solve the above model by the use of Gaussian Bayesian Derivative Classifier based on the characteristics of the relationship between microblog users and the behavior characteristics of users themselves. Finally, the key factors affecting the evaluation model are deeply studied by simulation experiments. The structure of this paper is described as follows: Section 1 describes the research status of microblog users' influence; Section 2 presents the evaluation index and model of user's influence; solution is made based on Gaussian Bayesian Derivative Classifier in Section 3; simulation experiments are conducted in Section 4; Section 5 summarizes the whole paper.

2 Influence Evaluation Model of Microblog User

As a convenient social platform, microblog plays an important role in the information dissemination. Since there are many active users on microblog, and the ways of information release are wide, and the information dissemination is featured by rapid speed and wide range, which are not conducive to the information management on microblog, and extremely easy to create public opinion on the Internet, this paper will propose a system to evaluate the influence of microblog user and thus calculate the influence index of users based on actual data for the convenience of managing the microblog social platform.

User influence is an index to measure the dissemination capacity of microblog users, and the greater the user's influence is, the greater the dissemination capacity and the impact on individual and even the society. User's influence factors include the relationship between the user's followers and fans and all kinds of behavior of users, which will directly determine the user's influence, and the influence evaluation system is constituted by analyzing the user's influence factors. This paper takes the coverage rate H of actual influence person-time of information dissemination, user's activity and connection degrees as the evaluation indexes for microblog influence. Coverage rate includes the number of fans and followers, and the activity degree includes the number of user's original tweets, the frequency of retweeting and commenting and the number of private letters, and the dissemination capacity includes the frequency of forwarding, reading, collecting, liking and commenting original tweets.

Specific algorithm of influence evaluation of microblog user is as follows:

1. The data are initialized. N users with higher activity degree and greater connection degree are randomly selected from all the microblog users, and the microblog data of these N users are collected, and the collected data include the number of original tweets, fans and followers, the frequency of being tweeted and tweeting, being commented and commenting, being collected and collecting, etc.
2. The user' activity degree is calculated. The user's activity degree can affect the user's influence

to some extent, and there are a lot of silent fans on microblog who have followed some users, but they cannot help the dissemination of information and thus they are unhelpful for the user's influence. As a result, silent fans are removed through multi-user activity degree, and the activity degree is calculated as follows:

$$T(i) = \frac{\omega_1 N + \omega_2 F + \omega_3 L}{t} \tag{1}$$

Where T(i) is the activity degree of the user i, and N is the number of original microblogs of the user i, and F is the number of fans of the user i, and L is the number of followers of the user i, and ω1,ω2,ω3 are weighted values of corresponding impression factors respectively, and t is the time interval of a time period.

3. The user's connection degree is calculated. Every microblog user is connected to each other, and the higher the connection degree is, the greater the information dissemination capacity, which will also enhance the influence of users. The calculation of connection degree between users is similar to that of user's activity degree, and the calculation is as follows:

$$T(i, j) = \frac{\lambda_1 U + \lambda_2 V + \lambda_3 S}{t} \tag{2}$$

Where T(i,j) is the connection degree between the user i and the user j, and U is the number of tweets of the user j forwarded by the user i, and V is the number of tweets of the user j commented by the user i, and S is the number of tweets of the user j collected by the user i, and λ1,λ2,λ3 are weighted values of corresponding influence factors respectively, and t is the time interval of a time period.

4. The user's coverage degree is calculated. The user's coverage degree is the number of active microblog fans of the user i, which has a great influence on the dissemination capacity of microblog and thus can indirectly improve the influence ability of microblog users. The calculation formula is as follows:

$$H(i) = \frac{\sum \mathrm{R} \cup \mathrm{V} \cup M}{N} \tag{3}$$

Where R, V and M are the crowd covered when counting the number of retweeting, commenting and following of the user, and N is the number of nodes on the Internet.

5. The influence of microblog user is calculated. The influence model proposed in this paper is

composed of three indexes, namely, user's activity, connection and coverage degrees, and its calculation formula is as follows:

$$X = \alpha T(i) + \beta T(i, j) + \gamma H(i) \tag{4}$$

Where T(i) is the user's activity degree, and T(i,j) is the connection degree between the user i and the user j, and H(i) is the user's coverage degree, and α, β and γ are weighted values of three influence factors.

6. It's required to repeat step 2 to step 5 until the influence of all users is calculated, and then jump to step 7.
7. The user's influence is sorted. Sorting the user's influence can find out the users with high influence in a more intuitive manner.

3 Solution Method Based on Gaussian Bayesian Derivative Classifier

In order to reduce the impact of zombie fans and spam microblogs on evaluation results, this paper proposes a model to evaluate the influence of microblog users based on Gaussian Bayesian Derivative Classifier, which considers both the relationship characteristics between microblog users and the behavior characteristics of each user, and Naive Bayesian Classifier with continuous attribute is established to identify the zombie fans, which can improve the classification efficiency and reliability to a certain extent. Specific algorithm is as follows:

Step1: The data are initialized. N users with higher activity degree and greater connection degree are randomly selected from all the microblog users, and the microblog data of these N users are collected, and the collected data include the number of original tweets, fans and followers, the frequency of being tweeted and tweeting, being commented and commenting, being collected and collecting, etc.

Step2: X_1，…，X_n are taken. C is the continuous microblog attribute (number of original tweets, fans and followers, the frequency of being tweeting and tweeting, being commented and commenting, being collected and collecting, etc.) and category. x_1，…，x_n, C is the value, and D is the selected data set with N records, and the data are generated randomly from the mixed distribution P, in which x_{ij}($1 \le i \le n$，$1 \le j \le N$) and c_j are the j^{th} recorded observation of X_i and C in data set D.

Step3: The users are classified by Gaussian Bayesian Derivative Classifier to remove zombie users. Assuming that G_1,G_2 are two k-dimension populations, in which G_1 is the real user and G_2 is the zombie user, and its distribution density is $p_1(u)$,$p_2(u)$ respectively, and user group $u=(u_1,u_2,\ldots,u_m)$, in which the

probability of u_1 coming from G_1 is q1, and the probability of u_2 coming from G_2 is (1-q1). The k-dimension space R^k is divided into (R_1,R_2), which satisfies:

$$R^k = \left\{ R_1, R_2 \mid R_1 \cup R_2 = R^k, R_1 \cap R_2 \right\} \tag{5}$$

If $u \in R_1$, it means that the user u_1 comes from the real user group G_1; if $u \in R_2$, it means that the user u_1 comes from the zombie user group G_2. No classification method can meet 100% accuracy rate, and the probability that the real user group G_1 is misjudged as the zombie user group G_2 based on this classification is calculated as per equation (6), and the probability that the zombie user group G_2 is misjudged as the real user group G_1 is calculated as per equation (7).

$$p(2|1,R) = \int_{R_2} p_1(u_i)dx \tag{6}$$

$$p(1|2,R) = \int_{R_1} p_2(u_i)dx \tag{7}$$

Then, average classification error f(R_1,R_2) of classification results is calculated as follows:

$$f(R_1,R_2) = q_1 c(2|1) P(2|1,R) + q_2 c(1|2) P(1|2,R) \tag{8}$$

Step4: The similarity between the number of fans F_i of selected user and the number of real fans $F_i^{'}$ is measured, and the calculation is as follows:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{F_i - \overline{F}}{S_{F_i}} \right) \cdot \left(\frac{F_i^{'} - \overline{F^{'}}}{S_{F_i^{'}}} \right) \tag{9}$$

$$\overline{F} = \frac{1}{n} \sum_{i=1}^{n} F_i \tag{10}$$

$$\overline{F^{'}} = \frac{1}{n} \sum_{i=1}^{n} F_i^{'} \tag{11}$$

$$S_{F_i} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (F_i - \overline{F_i})^2} \tag{12}$$

$$S_{F_i'} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(F_i' - \overline{F_i'})^2} \tag{13}$$

Where $\overline{F}$ is the average number of fans of an user, and $\overline{F_i}$ is the average number of real fans of an user, and S_{F_i} , $S_{F_i'}$ are the average number of fans and real fans of an user respectively. In this way, the list of fans can be truly restored, and zombie fans are removed from the list of fans of the user, and then the data after the removal of zombie fans can be recalculated.

Step5: The user's activity, connection and coverage degrees are calculated as per equations (1), (2) and (3) respectively.

Step6: The influence of each user is calculated as per equation (4).

Step7: The influence of each user is sorted from the largest to the smallest.

4 Simulation Experiment

This paper counts such data as basic user information, list of fans and tweets posted by the user within a month by taking the users of Sina microblog as experimental data, and the user's influence ranking calculated by the model proposed in this paper and the HRank model is compared with the influence ranking of existing users of Sina microblog, and statistics are made on top 10 user influence ranking of each model, and the results obtained are shown in Table 1.

Ranking	Sina Microblog	HRank	Model in This Paper
1	People's Daily	2	1
2	Global Times	1	2
3	Chutian City Newspaper	3	3
4	China Youth News	4	4
5	Xinmin Evening News	7	6
6	Beijing News	6	5
7	Yangcheng Evening News	8	8
8	Morning Post	5	7
9	Peninsula City News	9	9
10	Dahe Daily	10	10

Table 1: Top 10 in Media/Newspaper Influence of Each Model

Table 1 shows top 10 users in the user's influence ranking of each model obtained after counting

and calculating the newspaper microblog users in the microblog media ranking. It can be seen from Table 1 that the users who have a large number of fans are not necessarily influential, and that the user's influence ranking calculated by the model proposed in this paper is basically the same as that of Sina microblog. It thus can be seen that the model proposed in this paper is feasible.

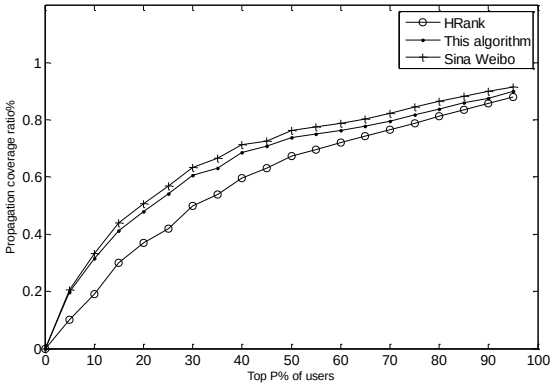

Fig. 1 Comparison of Coverage Rates of Different Algorithms

Fig. 1 shows the comparison of coverage rates of influence person-time of top p% users under different algorithms, and considers the advantage-disadvantage relationship of an algorithm, and the results are shown in Fig. 1. As can be seen from Fig. 1, the algorithm proposed in this paper is slightly better than HRank algorithm, and its ranking is slightly lower than the real ranking on Sina microblog, but it is closer to the real ranking, which indicates that the algorithm proposed in this paper is very reasonable in the classification and screening of microblog information.

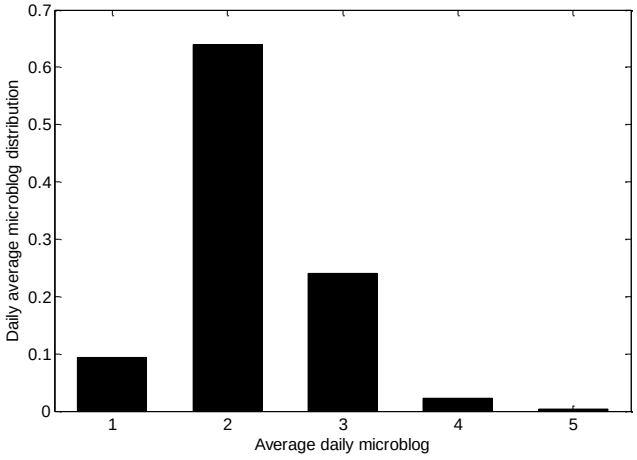

Fig. 2 Distribution Diagram of User's Average Daily Tweets

Fig. 2 shows the distribution of user's average daily tweets, and counting the user's average daily tweets can remove the inactive or zombie users from these users in a more intuitive manner. As can be seen from Fig. 2, most users have 2-3 tweets per day, so an user with less than 1 tweet per day or only one tweet per day can be regarded as inactive or zombie user, and such users can be ignored in the evaluation of user's influence.

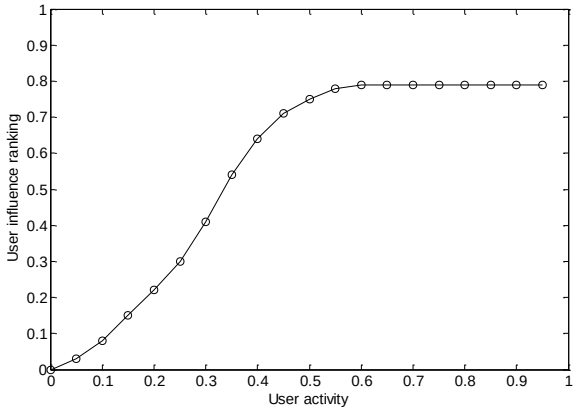

Fig. 3 Relationship between User's Activity Degree and Influence

As mentioned above, the user's activity degree is related to the ranking of user's influence, and the relationship between the user's activity degree and influence is studied in this paper, and the results are shown in Fig. 3. As can be seen from Fig. 3, there is almost a linear relationship between the user's activity degree and influence ranking within a certain range, but after exceeding a certain range, the user's activity degree has almost no significant impact on the user's influence ranking, which indicates that the user's influence ranking is also affected by other factors.

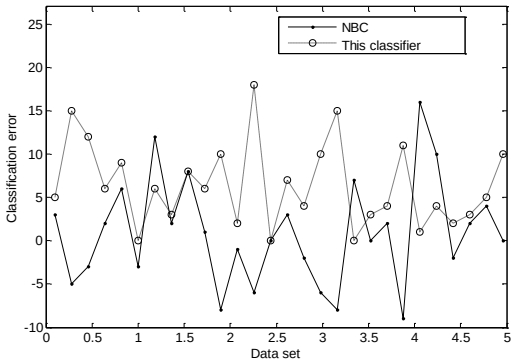

Fig. 4 Comparison of Classification Errors between NBC and Algorithm Proposed in This Paper

Five sets of data are classified and tested in this paper, and the error comparison is carried out with NBC algorithm, and the results are shown in Fig. 4. As can be seen from Fig. 4, the error curve chart of the algorithm proposed in this paper is always above the zero line, which indicates that the classification results of the algorithm proposed in this paper are better than NBC algorithm, and the curve difference of the algorithm proposed in this paper is smaller than NBC algorithm, which indicates that the algorithm proposed in this paper tends to be more stable.

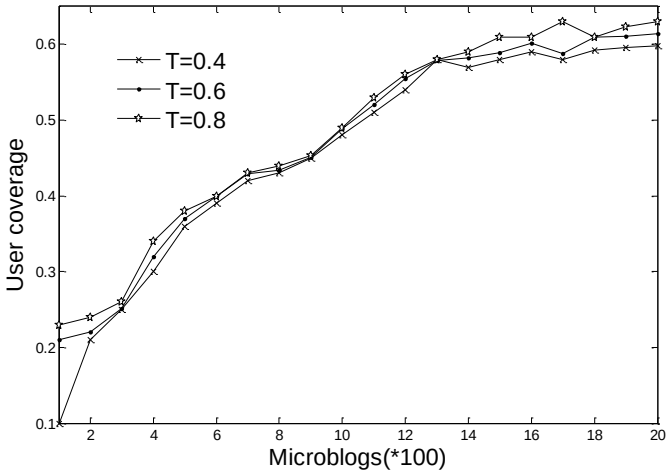

Fig. 5 Comparison of User's Coverage Degree and Number of Tweets at Different Activity Degrees

Finally, the influence factors of the user's coverage degree are discussed, which may be a positive correlation with the number of tweets according to the calculation formula, and the experiment environment can be changed to compare whether different activities have an impact on the calculation results. As can be seen from Fig. 5, the greater the activity degree is, the wider the user's coverage degree. Therefore, the user's activity degree will have an impact on the user's coverage degree. The number of active microblog fans of an user will be affected by whether an user updates the microblog, which directly reflects the user's influence. When there are less tweets, the influence of the activity degree is obvious, and when there are more tweets, the activity degree will also have a greater impact on the coverage degree.

5 Conclusion

As for the evaluation problem of microblog user's influence, this paper proposes an evaluation mode based on Gaussian Bayesian Derivative Classifier, which first presents a model to depict the influence

of microblog user by combining such indexes as activity, connection and coverage degrees, etc. and then gives the solution method of the above model using Gaussian Bayesian Derivative Classifier based on the characteristics of the relationship between microblog users and the behavior characteristics of users themselves. Finally, the key factors affecting the evaluation model are deeply studied by simulation experiments by taking Sina microblog users as experimental data. The results show that this algorithm has better adaptability than NBC algorithm. In the follow-up study, the dissemination characteristics of microblog information can be considered to improve the evaluation model of user's influence.

References

[1] Faliang H, Shi H, Dalin W, ge Y.(2016). Mining Topic Sentiment in Microblogging Based on Multi-feature Fusion . *Chinese Journal of Computers*(pp. 872-888).

[2] Bingyu L, Cuirong W, Cong W, Junwei W, Xingwei W.(2017). Microblog Community Discovery Algorithm Based on Dynamic Topic Model with Multidimensional Data Fusion. *Journal of Software*(pp.246-261).

[3] Yangsen Z, Jia Z, Anjie T.(2017). A Quantitative Evaluation Method of Micro-blog User Authority Based on Multi-Feature Fusion. *Acta Electronica Sinica*(pp.2800-2809).

[4] Xiao S, Xiaoqi P, Min H, Fuji R.(2017). Extended Multi-modality Features and Deep Learning Based Microblog Short Text Sentiment Analysis. *Journal of Electronics & Information Technology*(pp.2048-2055).

[5] Yang L, Yiheng C, Ting L.(2016). Survey on Predicting Information Propagation in Microblogs. *Journal of Software*(pp.247-263).

[6] Cha M, Haddadi H, Benevenuto F, Gummadi K P.(2010).Measuring user influence in twitter: The million follower fallacy. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, Menlo Park: AAAI Press*(pp.10-17).

[7] Wentao X, Feng L, Erzhou Z.(2016).Research on Novel Ranking Algorithm of Microblog User's Influence Based on Map Reduce. *Computer Science*(pp.66-70).

[8] Shaohua W, Xiaojuan M, Yong H. (2015).Impact evaluation of microbloggers based on improved PageRank algorithm. *Journal of Sichuan University(Natural Science Edition)*(pp.1040-1044).

[9] Weiguo Y, Yun L.(2015). Growth Law of User Characteristics in Microblog. *Journal of Computer Research and Development*(pp.522-532).

[10] Jia-Xin M, Qiqun L, Min Z, Shaoping M.(2014). Social aInfluence Analysis for Micro-Blog User Based on User Behavior. *Chinese Journal of Computers*(pp.791-800).

[11] Yaqi W, Jing W, Haibin Y.(2014). An evolution model of microblog user relationship networks

based on complex network theory. *Acta Physica Sinica*(pp.208902-208902).
[12] Zhiming X, Dong L, Ting L, Sheng L, Gang W, Shulun Y. (2014). Measuring Similarity between Microblog User and Its Application . *Chinese Journal of Computers*(pp.207-218).
[13] Haoran L, Meiting S, Lei L, Yongji L, Bin L.(2017). Study on Bayesian network structure learning algorithm based on ant colony node order optimization. *Chinese Journal of Scientific Instrument*(pp.143-150).
[14] Guofeng Z, Yan Y, Zhurong Z, Zhongyun Y, Fengjiao H.(2014). A Method of Calculating the Influence of Micro-Blog User Based on Domain. *Journal of Southwest University*(pp.145-151).
[15] Hao G, Yuliang L, Yu W, Liang Z.(2012). Measuring user influence of a microblog based on information diffusion. *Journal of Shandong University (Natural Science)*(pp.78-83).
[16] Shaowu Z, Jie Y, Hongfei L, Xiaohui W. (2015).A Micro-bolog User Influential Model Based on User Analsis. *Journal of Chinese Information Processing*(pp.59-66).
[17] Shuangcheng W, Ruijie D, Ying L.(2012). The Learning and Optimization of Full Bayes Classifiers with Continuous Attributes. *Chinese Journal of Computers*(pp.2129-2138).
[18] Bingwu F, Zhiqiu H, Yong L, Yong W.(2016). Quantitative Analysis Method of Dynamic Fault Tree of Complex System Using Bayesian Network. *Acta Electronica Sinica*(pp.1234-1239).
[19] Yuhu C, Yaoyao T, Xuesong W.(2011). A Selective Bayesian Classifier Based on Change of Class Relevance Influence. *Acta Electronica Sinica*(pp.1628-1633).
[20] Shuangcheng W, Rui G, Ruijie D.(2017). Learning and optimization of dynamic naive Bayesian classifiers for small time series. *Control and Decision*(pp.163-166).