Identification of Micro-blog Opinion Leaders based on User Features and Outbreak Nodes

https://doi.org/10.3991/ijet.v12i01.6139

Lin Cui
Nanjing University of Aeronautics and Astronautics, Nanjing China
Suzhou University, Suzhou, China
jsjxcuilin@ nuaa.edu.cn

Dechang Pi
Nanjing University of Aeronautics and Astronautics, Nanjing China
nuaacss@126.com

Abstract—At present, recognition of micro-blog opinion leaders mainly depends on the number of users posting micro-blogs, registration time, the number of good friends and other static attributes. However, it is very difficult to obtain the ideal recognition results through the above mentioned methods. This paper puts forward a new method that identifies the opinion leaders according to the change of user features and outbreak nodes. Deeply analyzing various attributes and behaviors of users, on the basis of user features and outbreak nodes, user's attribute features are regarded as the input variables, behavior features of the user and outbreak nodes are regarded as observed variables. The probability as an opinion leader is the latent variable between input variables and observation variables, and the constructed probability model is used to recognize micro-blog opinion leaders. Experiments are carried out on the two real-world datasets from Sina micro-blog and Twitter, and the comparative experimental results show that the proposed model can more precisely find the micro-blog opinion leaders.

Keywords—Micro-blog, Opinion leaders, User features, Influence strength, Outbreak nodes

1 Introduction

With the rapid development of Internet, especially Web 2.0 technology, micro-blog quickly gathers a large number of users and plays more important social roles with its high efficiency and convenience. Micro-blog has become one of important platforms in the domain of network public opinion because of its real-time. Hence, how to identify opinion leaders of micro-blog is a very challenging research problem.

Micro-blog opinion leaders refer to the users who have a larger number of fans, their active degrees are higher and they often publish original posts. The published posts from opinion leaders can produce enormous spread diffusion effects. We deeply study how to mine micro-blog opinion leaders combining with outbreak node discovery and user features and analyze the micro-blog network public opinions by some experiments. User attribute features are regarded as the input variables for identifying opinion leaders, behavior features and outbreak nodes are regarded as observed variables to distinguish whether the user is an opinion leader or not. The probability as an opinion leader is the latent variable between input variables and observation variables.

In our experiments on two real-world datasets, the proposed micro-blog opinion leaders identification model based on user features and outbreak nodes outperforms state-of-the-art algorithms including the traditional Bayesian algorithm and SVM algorithm. To summarize, the main contributions of this paper are three-fold:

- We propose a probabilistic generative graph model for identifying micro-blog opinion leaders, which encodes both user features and outbreak nodes defined by us.
- We stipulate the user attribute features to be the input variables, behavior features of the user and outbreak nodes to be observed variables. The probability as an opinion leader is the latent variable between input variables and observation variables.
- We conduct extensive experiments to evaluate the proposed method on two real-word datasets collected from Sina micro-blog and Twitter, respectively. Experimental results show that our proposed method outperforms state-of-the-art methods for identifying opinion leaders.

The rest of the paper is organized as follows: we review previous work in Section II, and list some preliminaries in detail in Section III. Sections IV introduce the micro-blog opinion leaders identification model based on user features and outbreak nodes and section V depicts the relative framework. The experimental analysis results are presented in Section VI, followed by the conclusion and future work in Section VII.

2 Related Work

The conception of the opinion leader was first defined by America communication scholars Lazarsfeld [1]. He proposed that, the opinion leader was someone who played some influential mediating role in the social groups during the passing process of media information, and also referred to two stage flow propagation model. In 1960's and 1970's, theories on opinion leaders have become one of the most important researches in behavior methodology and media sociology.

Opinion leaders have already been studied in different fields, for example, references [2] and [3] studied the application of opinion leaders in the information diffusion. References [4] and [5] emphasized the opinion leaders in marketing management.

Entering the twentieth century, the development of Web 2.0 technology promoted the appearance of many network communities. People could exchange information and communication ideas in different communities. There also exist opinion leaders in online communities and these opinion leaders affect the purchase behavior, political views of the people to a certain extent [6]. Accordingly, many scholars began to study the influence power of opinion leaders under the platform of social networks. Reference [7] studied the opinion leaders under the micro-blog community. Reference [8] considered opinion leaders under the online forum, analyzed the features that opinion leaders had and the difference between the posts published by opinion leaders and the posts published by other users.

In recent years, Twitter as the most popular micro-blog service has become the one of the research hot spots focused by scholars. References [9] and [10] adopted the number of fans and micro-blog forwarding number to measure the user influence power in Twitter.

3 Preliminaries

3.1 Definitions of User Feature Variables

Through intuitively analyzing micro-blog, it can be found that identifying opinion leaders needs to consider user attributes and user behaviors. User attributes refer to the related features of user attribute information, such as the nicknames of the user, the number of fans and the number of friends mainly reflecting the user’s credibility and determining the user's probability of being an opinion leader. Behavior characteristics refer to the behavior habit of users to publish micro-blog, for example, whether the released micro-blog belongs to the forwarding micro-blog, whether the released micro-blog is @ other users, and when forwarding micro-blog, whether comments are executed and so on.

This paper defines the following four kinds of user attributes. The first is user’s categories. Based on the existing categories of micro-blog users, this paper divides micro-blog users into ordinary individual users, special individual users, ordinary enterprise users and special enterprise users, among which, each category is set to different weight value according to its importance. The second is user’s activeness degree, which is used to measure the frequency that users post micro-blogs. The higher the user activeness degree is, the bigger the frequency that the user utilizes micro-blog and the reliability is also higher. The third is the value about fans, which reflects the quality of fans, the higher the value is, the better the quality of fans and the credibility is also better. The fourth is the value of good friends, which is obtained by defining the class set of users, the total number of friends focusing on some users, the total number of zombie friends and the number of good friends belonging to the user classification.

This paper also defines five kinds of user behavior characteristics. The first one is micro-blog original ratio, which measures the behavior tendency that the user issues micro-blogs. The low original ratio indicates that the user tends to transfer micro-blog and less publishes the original micro-blog. The second one is non-empty forwarding ratio, which measures whether comments are released when the user forwards micro-blog, if not publishing comments, and then non-empty forwarding ratio is empty. The third one is the original micro-blog interaction, which measures the interactivity between a user and other users, because micro-blog marketing activities contain a large number of information contents such as "forwarding and @ a friend", "@ some person" and so on. In order to attract customers to participate in micro-blog, we only consider the original micro-blog interactivity released by users. The fourth one is non marketing activity participation, which measures the frequency that the user participates in the micro-blog marketing activity. The fifth one is URL usage rate, which is a kind of behavior extent that measures the user to guide other users linking to the specified address.

3.2 Mining Outbreak Nodes

Outbreak nodes are able to significantly increase the spread scope during the posts dissemination, namely after outbreak nodes forwarding micro-blog posts, there would be a large number of users to follow to participate in forwarding the posts. Suppose $P(m_r) = \{p_1, p_2, \mathrm{L}\ \ , p_N\}$ is micro-blog posts set for analyzing, m_r represents the outbreak node. We also define the outbreak index value $\phi(m_r)$ to measure the importance of outbreak nodes as below:

$$\phi(m_r) = \sum_{i=1}^{N}\left[\frac{num(m_r,b_i) - num(m_r,m_e)}{num(m_r,b_i)}\right]\Big/ N \qquad (1)$$

where the calculation of outbreak nodes relates to the shortest distance between the outbreak node m_r and the source node b_i. The closer the distance between outbreak node m_r and the source node b_i is, the more quickly micro-blog can be spreading to the outbreak node m_r through fewer forwarding times. $num(m_r, b_i)$ denotes the number of edges between all outbreak nodes m_r and the source node b_i, $num(m_r, m_e)$ represents the number of edges between the outbreak node m_r and the outbreak node m_e. Through different values of outbreak nodes, outbreak nodes can be distinguished. The greater the value of an outbreak node is, the more significant its explosive effect is.

3.3 Latent Variable Models

This paper constructs a model based on the recognition probability of opinion leaders. In our constructed model, user attributes are regarded as the input conditions, behavior characteristics and outbreak nodes are regarded as the performance metrics to test whether the user is an opinion leader or not.

Suppose that $X = \{x_1, x_2, \cdots, x_n\}$ represent n attributes set of the user, $Y^{(n)} = \{y_1, y_2, \cdots, y_n\}$ denote n kinds of user behavior characteristics, $Y^{(m)} = \{y_{n+1}, y_{n+2}, \cdots, y_m\}$ denotes the discovered outbreak nodes, *Z* denotes the probability as an opinion leader. The discrimination model of *P(Z|X)* represents the conditional probability of user x being an opinion leader. The generative model of *P(Y|Z)* represents the behavior of the opinion leader using micro-blog. The joint distribution model can be formulated as follows:

$$\begin{aligned} &P(z, y_1, y_2, \cdots y_n, y_{n+1}, y_{n+2}, \cdots y_m \mid x_1, x_2, \cdots, x_n) \\ &= P(z \mid x_1, x_2, \cdots, x_n) \prod_{t=1}^{m} P(y_t \mid z) \end{aligned} \tag{2}$$

4 Micro-Blog Opinions Leader Discovery Model Based On User Features And Outbreak Nodes

Subsequently, we define a discriminative model *P(O|X)*, which denotes the probability that the user is an opinion leader and is described as follows with Gauss distribution :

$$P(o \mid x_1, x_2, \cdots, x_n) = M(w^T a(x_1, x_2, \cdots, x_n), v) \tag{3}$$

where a is the user attribute vector, w is the weight vector, v is the Gauss model variables and its value is 0.5. As user behavior feature is a continuous variable, it is difficult to describe the distribution generation model *P(Y|Z)* by using simple linear distribution. In order to guarantee the accuracy of the model and avoid the complicated calculation, two-order behavior variables are defined for the corresponding behavior features, which judge whether to publish the original micro-blog. If there is not an empty forwarding micro-blog behavior, *y=1* otherwise *y=0*; if the original micro-blog @ other users, *y=1* otherwise *y=0*; and if using URL behavior, *y=1* otherwise *y=0*. Logistic distribution characterization is used to describe *P(Y|Z)* model. Giving the probability of being an opinion leader *Z* and its behavior characteristic, the probability corresponding to two-order behavior variables is shown as:

$$\begin{aligned} &P(y_{t1}{}^{(n)}=1\,or\;y_{t2}{}^{(m)}=1\,|\,z^{(i)},a_t{}^{(i)}) \\ &\quad=\frac{1}{1+e^{-(\theta_{t1}z^{(i)}+\theta_{t2}a_t{}^{(i)}+v)}} \\ &S.T.\;\; t1\in\{1,2,\mathrm{L}\;\;n\},\,t2\in\{n,n+1,\mathrm{L}\;\;m\} \\ &\qquad t=t1\cup t2 \end{aligned} \tag{4}$$

Suppose $\theta_t=[\theta_{t1},\theta_{t2}]^T,u_t{}^{(i)}=[z^{(i)},a_t{}^{(i)}]^T$, the formula (4) is denoted as :

$$\begin{aligned} &P(y_{t1}{}^{(n)}=1\,or\;y_{t2}{}^{(m)}=1\,|\,u_t{}^{(i)}) \\ &\quad=\frac{1}{1+e^{-(\theta_t{}^T u_t{}^{(i)}+v)}} \end{aligned} \tag{5}$$

By using the maximum likelihood estimation, the log likelihood function of the formula is solved; and by using the above formulas, the posterior probability that new user is an opinion leader can be also predicted.

5 The Proposed Framework

Combined with user attribute features, the behavior characteristic and outbreak nodes, this paper proposes an research framework of discovering opinion leaders, which is composed of the data preprocessing layer, topic discovery layer and opinion leaders discovery layer, as shown in Figure 1.

Each layer is introduced as follows respectively:

5.1 Data preprocessing layer

Data preprocessing layer includes some sub-steps, such as capturing micro-blog resources from Sina micro-blog and Twitter website, then executing text segmentation, removing stop words and text feature selection and so on.

5.2 Topic Discovery Layer

After getting the data from the data preprocessing layer, removal of words with no practical significances is executed and the processed information is obtained, namely the tag set and resource text. Then, latent topic comment resources are obtained by using LDA model and Gibbs sampling algorithm, and the tag clustering based on topic labels is performed and assigned to different topics. This layer can solve the defects of the limited fields of good opinion leaders, and lay the foundation for the discovery of opinion leaders.

5.3 Opinion Leaders Discovery Layer

Based on the discovered latent topics and aiming at the user's attribute features and behavior characteristics, outbreak nodes are firstly discovered. Then according to the proposed opinion leaders discovery model, opinion network relation matrixes are firstly constructed. Through the UCINET analysis tool, the index values on the relation matrixes are got, and the selected maximums are used to determine the weight of each index method. The comprehensive index values are sorted and the corresponding topic opinion leaders are found at last.

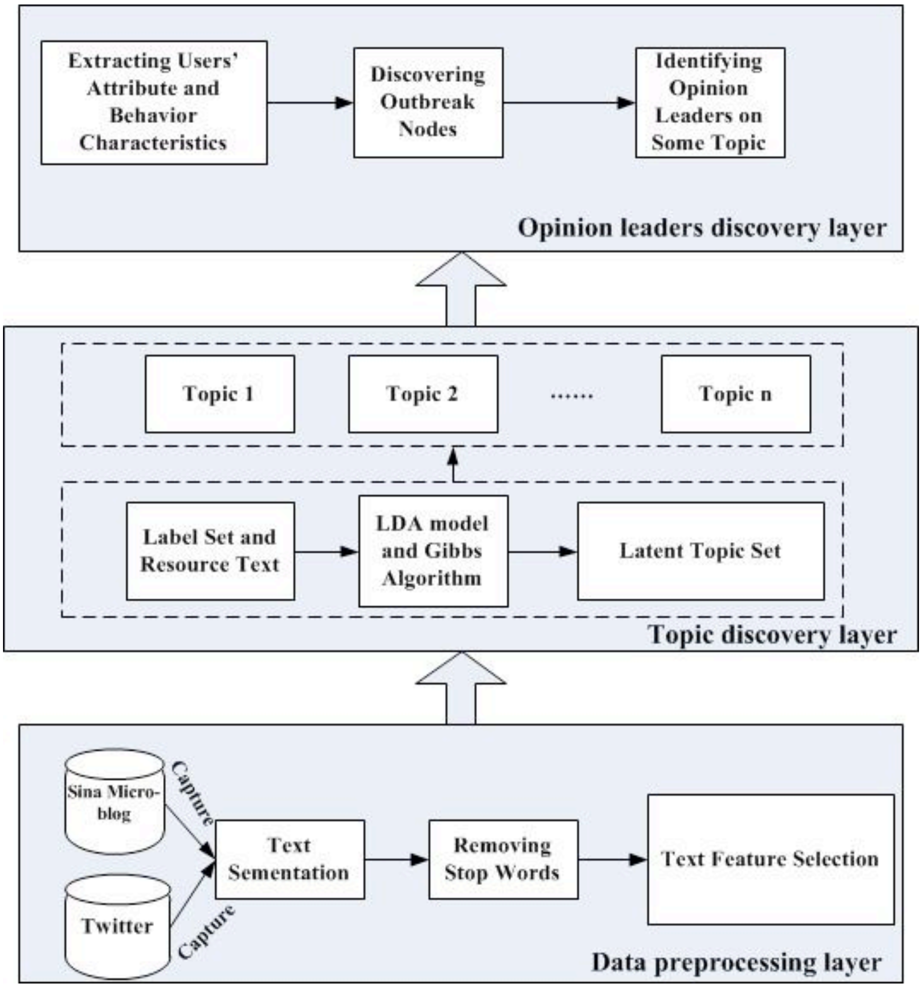
Extracting Users' Attribute and Behavior Characteristics
Discovering Outbreak Nodes
Identifying Opinion Leaders on Some Topic
Opinion leaders discovery layer
Topic 1
Topic 2
......
Topic n
Label Set and Resource Text
LDA model and Gibbs Algorithm
Latent Topic Set
Topic discovery layer
Sina Micro-blog
Twitter
Capture
Capture
Text Sementation
Removing Stop Words
Text Feature Selection
Data preprocessing layer

Fig. 1. Framework of micro-blog opinion leaders identification model based on user features and outbreak nodes

6 Experimental Analysis

In this section, we conduct experiments to evaluate the effectiveness of the proposed opinions discovery framework. Specifically, we aim to answer the following two questions. (1) Can the proposed framework improve the performance of discovering the opinion leaders by incorporating user features and outbreak nodes? (2) Is the proposed framework able to boost the accuracy of discovering opinion leaders compared with other state-of-the-art baseline methods?

We begin by introducing datasets and experimental settings, then we investigate the capability of the proposed framework to answer the first question, and we compare our model with the state-of-art opinion leader discovery methods to answer the second question.

6.1 Experimental Dataset

In order to objectively evaluate the effect of opinion leaders discovery model based on user features and outbreak nodes which is abbreviated as UF+OND OLDM, this paper uses two datasets that are from Sina dataset and Twitter dataset. These two datasets are shown in Table1. Sina dataset is collected from Sina micro-blog open platform “http://overseas.weibo.com” and we firstly execute the artificial discrimination of opinion leaders. Twitter dataset comes from Twitter online website “https://twitter.com”, the discrimination results of opinion leaders come from the user annotation provided by the dataset. In order to better describe the differences of behavior characteristics between opinion leaders and the general users, we choose different behavioral characteristics which are shown in Table I.

Table 1. Statistics of the Two Real-world Datasets

Datasets	No. of users	No. of micro-blog	No. of fans	Leader/Non-leader
Twitter	8,563	47,035	2,542	1,061/ 7,502
Sina	9,987	56,276	3,975	1,482/ 8,505

6.2 Evaluation Metrics

Confusion matrix is a visual tool in supervised learning, mainly for comparing the classification results and the true information. Each row in the matrix represents the predicted category of real examples, and each column represents an instance of the class [11]. In the confusion matrix, every instance can be divided into one of four types which are *TP* (True Positive), *FP* (False Positive), *FN* (False Negative) and *TN* (True Negative) respectively, as shown in Figure 2. Wherein, positive represents prediction examples from the positive class and negative denotes the prediction examples from the negative class; true implies the correct prediction and false refers to the prediction error.

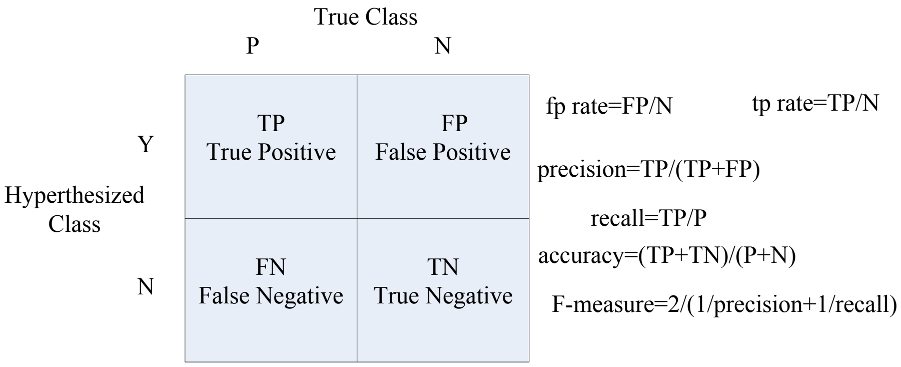

Fig. 2. Confusion matrix and common performance metrics

In order to evaluate the learning ability of the proposed model, *AUC* (Area under Curve) that is area under the *ROC* (Receiver Operating Characteristic) curve is adopted to evaluate a value of binary classifier [12]. *ROC* focuses on two indexes that are true positive rate *TPR=TP/[TP+FN]* and false positive rate *FPR=FP/[FP+TN]*. Intuitively, *TPR* represents the probability that positive cases can be judged to be right; *FPR* represents the probability that negative examples are misclassified as positive examples. In ROC space, the abscissa of each point is *FPR*, the vertical axis is *TPR*, *FPR* and *TPR* described the trade-off between *TP* (true positive cases) and *FP* (positive error) of the classifier. ROC curve passes through point (0, 0) and point (1, 1), usually at the top of the line $y = x$, so the value of AUC is between 0.5 and 1.

6.3 Experimental Result Analysis on the Proposed UF+OND OLDM Model

The purpose of learning the model UF+OND OLDM is to predict the opinion leaders, so we need to investigate the prediction ability of UF+OND OLDM model. We randomly select *60%* of the experimental dataset as training samples; the remaining data are regarded as test samples.

The following Figure 3 and Figure 4 denote the experimental results of the proposed UF+OND OLDM model on Sina dataset. Figure 3 is a *ROC* curve about the prediction results, which represents the best classification threshold value. Under this threshold value, true positive rate and false positive rate both have the most suitable value and *ROC* curve shows the best classification point is that the threshold is *0.54*. Figure 4 is a *P-R* curve under different threshold values, which shows the predictive ability of the proposed model on Sina dataset.

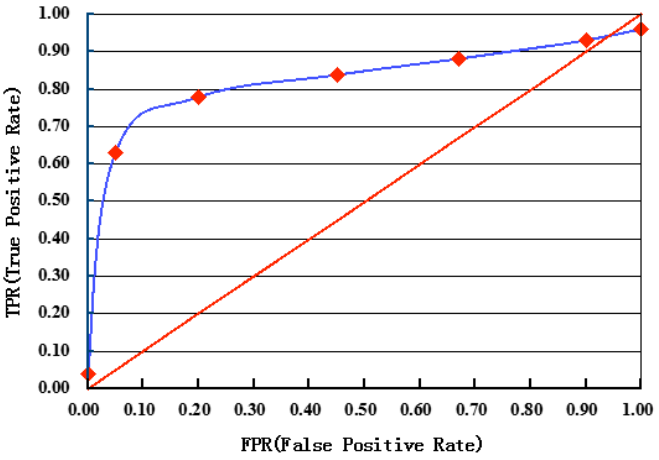

Fig. 3. *ROC* curve on Sina dataset

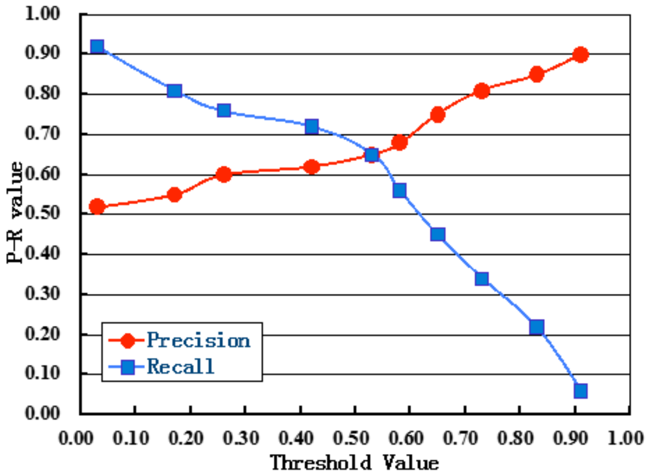

Fig. 4. *P-R* curve on Sina dataset

The following Figure 5 and Figure 6 denote the experimental results on Twitter dataset. Compared with Figure 3 and Figure 4, it can be seen that the model prediction results on *AUC* value on Twitter data set is better than the prediction results on Sina dataset. Under the circumstances that the model gets the optimization threshold, the prediction accuracies of Sina dataset and Twitter data set can both reach about *90%*, and the recall rate can also control in a high level. Hence, the prediction ability of the proposed UF+OND OLDM model is strong.

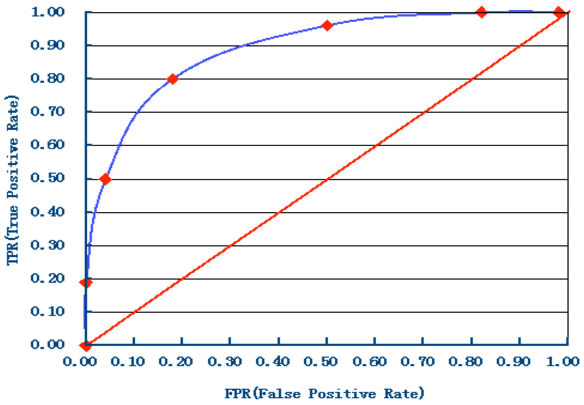

Fig. 5. *ROC* curve on Twitter dataset

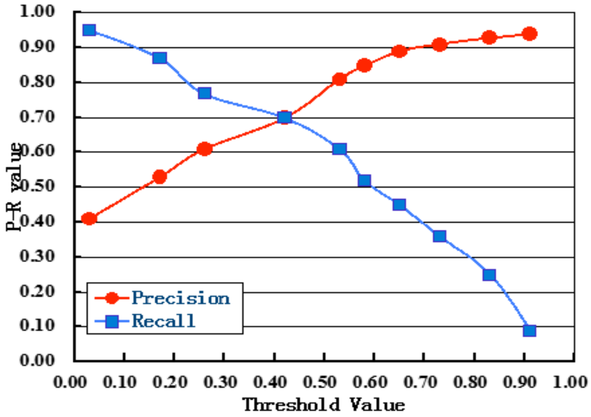

Fig. 6. *P-R* curve on Twitter dataset

6.4 Comparative Experiments between the Proposed Model and Other Baseline Methods

At present, aiming at the problems of recognizing opinion leaders, machine learning methods are widely used to classify opinion leaders. In order to evaluate the effect of the proposed UF+OND OLDM model, two baseline methods are adopted for the comparisons that are Bayesian algorithm [13] and SVM algorithm [14] respectively, among which, SVM method utilizes RBF kernel function. In all comparative experiments, we adopt ten-fold cross validation to estimate the relative parameters. The comparative experimental results are shown in Figure 7 and Figure 8. Figure 7 shows the comparisons of precision rate and recall rate on Sina dataset, Figure 8 presents the similar comparison results on Twitter dataset.

As Figure7 shows, on Sina data set, the precision rate of UF+OND OLDM model is higher about *8%* than that of Bayesian method and higher about *12%* than that of SVM method in classification learning. As Figure8 shows, on Twitter dataset, the precision rate and recall rate of UF+OND OLDM model are also higher than those of Bayesian method and SVM method. Combined with the experimental results of Sina dataset and Twitter data-set, it can be found that UF+OND OLDM model is better than Bayesian method and SVM method. In addition, before using the model or the classification algorithm, the time efficiency brought by the statistical characteristics is difficult to be calculated out. However, during the phase of the model learning and prediction, it can be found that the time efficiency of UF+OND OLDM model is almost equal with the time efficiency of traditional machine learning method, but in the accuracy and effectiveness, UF+OND OLDM model is higher than those of the traditional machine learning methods.

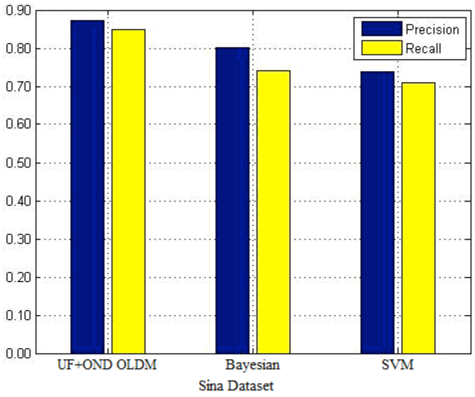

Fig. 7. Precision rate and recall rate on Sina dataset

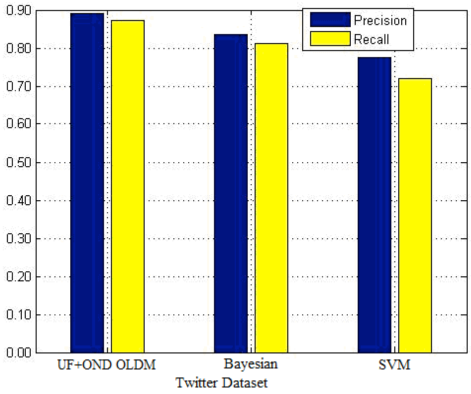

Fig. 8. Precision rate and recall rate on Twitter dataset

7 Conclusion And Future Work

With the popularly use of micro-blog and other social medias, the influence powers of opinion leaders under social network platforms are also growing. How to accurately and efficiently identify opinion leaders is a very challenging problem. This paper firstly deeply compares and analyzes the ordinary users and opinion leaders under the micro-blog platform, and then proposes that user attribute features are regarded as the input variables, user behavior features and outbreak nodes are regarded as observed variables, the probability as an opinion leader is the latent variable between input variables and observation variables. Comprehensively analyzing different experimental results on Sina dataset and Twitter dataset, it can be observed that the proposed UF+OND OLDM model has the better identification ability of opinion leaders compared with Bayesian method and SVM method. For the future work, it would be interesting to incorporate temporal factor [15, 16] into the proposed UF+OND OLDM model for identifying opinions leaders.

8 Acknowledgment

This work was supported by the Scientific Research Innovation Program for College Graduates of Jiangsu Province in China (NO.KYLX15_0321), the Key Project of Overseas Visiting for the Outstanding Young Backbone Talents at Universities and Colleges of Anhui Province in China (NO.gxfxZD2016267), the Key Project of Anhui Province Colleges and Universities Natural Science Foundation of China (No.KJ2016A768, No.KJ2014A250) and the Soft Science Research Project of Anhui Science and Technology Project (No.1607a0202071). Furthermore, thanks editors and reviewers for all the valuable advice.

9 References

1. P.F. Lazarsfield, “The people’s choice: how the votes makes up his mind in a presidential election,” New York: Colu. Uni. Press, 3rd edition, May 1968.
2. E. Bakshy, J.M. Hofman, W.A. Mason and D.J. Watts, “Everyone’s an influencer: quantifying influence on Twitter,” *Proc. of the 4th ACM Inter. Conf. on Web Search and Data Mining*, pp. 65-74, February 2011. https://doi.org/10.1145/1935826.1935845
3. M. Cha, H. Haddadi, F. Benevenuto and K.P. Gummad, “Measuring user influence in Twitter: the million follower fallacy,” *Proc. of the 4th Inter. AAAI Conf. on Web. and Social Media*, pp. 10-17, May 2010.
4. H.J. Li, Y. Tian, W.C. Lee, C.L. Giles and M.C. Chen, “Personalized feed recommendation service for social networks,” *Proc. of the Second Inter. Conf. on Social Com.*, pp. 96-103, August 2010. https://doi.org/10.1109/socialcom.2010.23
5. G. Karypis, “Evaluation of item-based top-n recommendation algorithms,” *Proc. of the Tenth Inter. Conf. on Infor. and Know. Man.* , pp. 267-274, November 2001. https://doi.org/10.1145/502585.502627
6. S.J. Yoon, “The antecedents and consequences of trust in online purchase decisions,” *Jour. of Inter. Mark.*, vol. 16, pp. 47-63, March 2002. https://doi.org/10.1002/dir.10008

7. J. Zhang, M.S. Ackerman and L. Adamic, "Expertise networks in online communities: structure and algorithms," *Proc. of the 16th Inter. World Wide Web Conf. Comm.,* pp. 221-230, May 2007. https://doi.org/10.1145/1242572.1242603
8. J. Wang, J.P. Zeng, B.H. Zhou and C.R. Wu, "Online forum opinion leaders discovering method based on clustering analysis," *Compu. Engi.*, vol. 37, pp. 44-46, May 2011.
9. A. Anagnostopoulos, R. Kumar and M. Mahdian, "Influence and correlation in social networks," *Proc. of the 14th ACM SIG. Inter. Conf. on Know. Disc. and Data Min.*, pp. 7-15, August 2008. https://doi.org/10.1145/1401890.1401897
10. R. Ghosh and K. Lerman, "Predicting influential users in online social networks," *Proc. of the 4th KDD Works. on Soc. Net. Ana.* , pp.215-224, May 2010.
11. J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," *Proc. of the 23rd inter. Conf. on Mach. Lear.*, pp.233-240, June 2006. https://doi.org/10.1145/1143844.1143874
12. T. Fawcett, "An introduction to ROC analysis," *Patt. Recog. Lett.*, vol. 27, pp.861-874, December 2005. https://doi.org/10.1016/j.patrec.2005.10.010
13. X.W. Yang, Y. Guo and Y. Liu, "Bayesian-Inference-based recommendation in online social networks," *IEEE Trans. on Para. and Dis. Sys.*, vol. 24, pp. 642–650, April 2013.
14. J.H. Fu, J.H. Chang, C.Y. Ke and S.L. Lee, "Confidence-based link/attribute inference based on friendship circles," *Proc. of the 12th Inter. Conf. on Adv. in Mob. Com. and Multi.*, pp.330-335, December 2014. https://doi.org/10.1145/2684103.2684113
15. Y. Lia, S. Q. Ma, Y.H. Zhang, R.H. Huang and Kinshuk, "An improved mix framework for opinion leader identification in online learning communities," *Know.-Bas. Sys.*, vol. 43, pp. 43–51, May 2013.
16. D. Zhou, I. Councill, H.Y. Zha and C.L. Giles, "Discovering temporal communities from social network documents," *Proc. of 7th IEEE Inter. Conf. on Data Min.*, pp.745-750, October 2006.

10 Authors

Lin Cui, she is studying for doctoral degree at the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics of China, Nanjing 210016, China, and she is also an associate professor at Intelligent Information Processing Lab, Suzhou University, Suzhou 234000, China. Her research interests include Web text mining and Social network. (e-mail: jsjxcuilin@ nuaa.edu.cn).

Dechang Pi, he is a professor at College of Computer Science and Technology in Nanjing University of Aeronautics and Astronautics of China, Nanjing 210016, China. His research interests include data mining and big data analysis. (e-mail: nuaacss@126.com).

Submitted, June, 23, 2016. Published as resubmitted by the authors on August, 11, 2016.