

Research Article

S3UCA: Soft-Margin Support Vector Machine-Based Social Network User Credibility Assessment Method

Kang Zhao , Ling Xing , and Honghai Wu 

School of Information Engineering, Henan University of Science and Technology, 471023 Luoyang, Henan, China

Correspondence should be addressed to Ling Xing; xingling_my@haust.edu.cn

Received 22 August 2021; Revised 26 September 2021; Accepted 10 November 2021; Published 25 November 2021

Academic Editor: Yugen Yi

Copyright © 2021 Kang Zhao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Among the algorithms used to assess user credibility in social networks, most of them quantify user information and then calculate the user credibility measure by linear summation. The algorithm above, however, ignores the aliasing of user credibility results under the linear summation dimension, resulting in a low evaluation accuracy. To solve this problem, we propose a user credibility evaluation method based on a soft-margin support-vector machine (SVM). This method transforms the user credibility evaluation dimension from a linear summation dimension to a plane coordinate dimension, which reduces the evaluation errors caused by user aliasing in the classification threshold interval. In the quantization of user information, the ladder assignment method is used to process the user text information and numeric information, and the weight assignment method of information entropy is used to calculate the weight assignment among different feature items, which reduces the errors caused by the inconsistency of the order of magnitude among different types of user information. Simulation results demonstrate the superiority of the proposed method in the user's credibility evaluation results.

1. Introduction

In the era of big data, the number of social network platforms and users has been growing exponentially, making social network platforms not only indispensable information interaction platforms and information communication media in people's daily lives but also huge and complex user groups [1, 2]. Users in social networks are important nodes for information dissemination on social platforms. The proliferation of malicious users (TNs) continues to challenge the smooth and healthy development of information dissemination on social platforms [3]. Meanwhile, the credibility assessment of social network users has important research significance in information screening, public opinion governance, network security, user identification, and other fields. Therefore, quantifying and evaluating the credibility of users in social networks has become an important research topic [4].

While social networks bring convenience to people's information exchange and emotional expression, their openness also leads to them being full of a large number of

TNs [5]. TNs generate a large number of false and malicious behaviors or information in social networks, and they increase their credibility within social networks by inventing profile information [6]. To ensure the accuracy and rationality of user credibility evaluation, it is necessary to strengthen the processing and quantification of each feature in the user profile information and user-generated content information, as well as improving the accuracy of user credibility evaluation algorithms [7–9]. To further study and solve these problems, researchers at home and abroad have proposed a variety of user credibility evaluation algorithm models and methods [10–15].

Machine learning algorithms are widely used in processing user information and quantitative evaluation [16], and researchers at home and abroad have made different research progress. Narayanan et al. [17] and Van Der Walt et al. [18] conducted quantitative research on the relevant profile information of both TNs and trusted users (TPs) in social networks through machine learning algorithms such as a decision tree algorithm and a PageRank algorithm. Slimi et al. [19], through the study of user-generated content

information, proposed to take the digital information and information distribution rule in user-generated content information as the quantitative object to measure the generated content information to achieve the evaluation of user credibility. To solve the problem of sparse information in user credibility evaluation, Verma et al. [20] proposed a new user credibility evaluation model, which extracts the relevant features of user profile information and user-generated content information, and they calculated the corresponding score by the linear sum of each feature to build the user credibility evaluation model using multitype information. In domestic studies, Sun et al. [21] comprehensively considered multiple evaluation indexes in user interaction information and user-generated content information. Based on the chain principle of a Bayesian network and a PageRank algorithm, they proposed for the first time a social network user evaluation model based on users' personal information and credibility transmission of adjacent users. Zhang et al. [22] put forward a kind of interactive message privacy protection scheme based on users' credibility, which is a scheme focused on the social behavior of network users that reflects their credibility. This study is the first to use the theory of the psychology of an item to assess the risk level of user interaction news to achieve credibility assessment to the user. Li et al. [23] identified and verified four new interaction attributes that affect the credibility of users in social networks: reply frequency, comment length, time difference, and domain similarity. Through logistic regression analysis of users' interaction information and generated content information from social networks, they studied the impact of interaction attributes in social networks on user credibility.

To explore and solve the problem of weight assignment and resource allocation between different information [24], Rahangdale and Thakar [25] not only assigned appropriate weights to user attribute characteristics such as user interaction behavior information on the network—including “mark,” “comment,” and “like”—to calculate credibility, they also considered the influence of mutual friends between users on user credibility. Thus, a method to evaluate user credibility in social networks is proposed, which avoids the impact of features with different importance on credibility assessment. Khan and Lee [26] proposed a model to evaluate user credibility based on multitype information of user profiles and user interaction information. This model is principally based on the user's spatial similarity. The model uses similar configuration files and interests to represent the similarity among user information such as the similarity between users' credibility. The model solves the weights between the user's credibility evaluation of the multitype information assignment problem. Zheng and Qu [27] used the entropy weight method to solve the weight assignment problem in four factors (social relationship strength, social influence scope, information value, and information transmission control) to propose a new user credibility evaluation model.

To solve the above problems, this paper first proposes a social network user credibility evaluation method model UCSSVM based on a soft-margin support-vector machine (SVM). The main contributions of this paper include: (1) the

step assignment method and information entropy weight distribution method are used to process the user profile information and user generated content information so as to solve the evaluation result error caused by the inconsistency of different types of user information. (2) The soft-margin SVM algorithm is used to solve the problem of user aliasing. (3) Based on the above methods, the user information is quantified, the measurement set of user credibility is calculated, and the measurement set of user credibility is evaluated in a two-dimensional plane to obtain the evaluation results. (4) Experiments on real user data sets verify the performance advantages of the proposed method.

2. Related Work

This paper first proposes a social network user credibility evaluation method model UCSSVM based on a soft-margin SVM. The model processes the user profile information and user generated content information using the ladder assignment method and information entropy weight distribution method and uses the soft-margin SVM algorithm to evaluate the measurement set of user reliability so as to avoid the evaluation result error caused by different types of user information in different orders of magnitude and types in other algorithms and solve the problem of user aliasing at the classification threshold in other algorithms.

2.1. Ladder Assignment Method. The ladder assignment method is used principally to solve the problem whereby the order of user information is not uniform in the calculation of user credibility. This method is designed to divide the digital information in the user information linearly and uniformly and to assign the values to different units linearly and continuously. The ladder assignment to the text of user information calculates the similarity between the feature items of the text and the tag types of the known tag set, identifying the type of the feature items of the text and assigning the value according to the different types of feature items.

The assignment of text data in the ladder assignment method is based on the Jaccard Similarity method to identify the text data states of user features to achieve the assignment of different states of user information feature items. The user text information feature set u and the known label set v are calculated, and the similarity degree $\text{InS}(u, v)$ between the two data sets is obtained:

$$\text{InS}(u, v) = \frac{\text{In}_u \cap \text{In}_v}{\text{In}_u \cup \text{In}_v} \quad (1)$$

The ladder assignment of digital information data is the uniform division of linear continuous distributed data, and the linear continuous assignment is carried out for different partition units. From the mathematical algebraic expression equation (2) of the step function, the ladder assignment function of the numeric information data can be obtained as equation (3).

$$f(x) = [Kx], \quad (2)$$

$$f(x) = \left\lfloor \frac{N}{M+1}x \right\rfloor, \quad (3)$$

where $K = (N/M + 1)$, the divergence of the numeric data of the user is $M = M_{\max} - M_{\min}$, the order of the ladder assignment is N , where N is an integer not equal to 0, and the value of the numeric data of the user is x . Therefore, the expression of ladder assignment of numeric data in user information features is :

$$f(x) = \begin{cases} \left\lfloor \frac{N}{M}x \right\rfloor, & M_{\min} < x < M_{\max}, \\ 0, & \text{other.} \end{cases} \quad (4)$$

2.2. Weight Distribution Method of Information Entropy. The weight distribution method of information entropy means that the smaller the information entropy of the computed feature term, the greater the information provided by the feature term, the greater the role it plays in the comprehensive evaluation, and the higher the weight.

In the evaluation set with N feature items, the weight values of different feature items in the evaluation set are calculated by the judgment matrix. The calculation equation of the evaluation set is

$$f(x) = n_1x_1 + n_2x_2 + \dots + n_nx_n, \quad (5)$$

where x_n represents the n th feature and n_n represents the weight of the n th feature.

The judgment matrix A_G of $n \times n$ can be obtained from the evaluation set. The eigenvector corresponding to the maximum eigenvalue of the judgment matrix A_G is calculated, and then the weight of n eigenterms is assigned.

$$A_G = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \dots & \dots & \ddots & \dots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{bmatrix}. \quad (6)$$

In the judgment matrix of the information entropy weight distribution method, e_n represents the entropy value of the n th feature in the evaluation set, $M_{\max} - M_{\min}$ is the divergence of the user information feature, and p_m represents the ratio of the entropy value of any two items in the evaluation set, as shown in equations (7) and (8).

$$e_n = M_{\max} - M_{\min} + 1, \quad (7)$$

$$p_{ij} = \frac{e_j}{e_i}. \quad (8)$$

By calculating the judgment matrix, the maximum eigenvalue of A_G was obtained as λ_{\max} , and the eigenvector of λ_{\max} was normalized to obtain the weight of each eigenterm in the evaluation set as: (n_1, n_2, \dots, n_n) .

2.3. Soft-margin SVM Algorithm. In the evaluation method of user credibility, the method of one-dimensional linear arrangement evaluation is intended to obtain the credibility value by the linear sum of the quantized feature items according to multiple information features of the user. There are some problems with this method, such as user aliasing at the classification threshold and immobilization of the weight assignment of each feature item.

The basic idea of an SVM [28, 29] is to change the original data space into a high-dimensional feature space through nonlinear changes and then to obtain the optimal linear classification surface in this new space so that the hyperplane can correctly separate the two types of samples and maximize the classification interval. The problem of user credibility aliasing at the classification threshold is avoided. The characteristic benefit of an SVM algorithm is that it finds the best compromise between the learning accuracy of a specific training sample in the model and offers the ability to identify any sample without errors according to the limited sample information, thereby obtaining the best robustness.

Retaining the order to solve the classification evaluation of user data within the hyperplane is inseparable. The article introduces the slack variable in constraint conditions. An interval hyperplane is a reasonable proportion of user data as it simultaneously avoids introducing slack variable credibility evaluation errors to the customer while adding the balance coefficient C in the objective function to solve this problem. Therefore, the equilibrium coefficient C is defined as the weight coefficient between the hyperplane with the largest interval of the balance support vector machine and the guarantee of the minimum deviation of the data points.

3. User Credibility Evaluation Method

3.1. User Credibility Assessment Framework. The social network user credibility evaluation method proposed in this paper includes four steps: (1) collect the data set required by the experiment through the API interface provided by the social network platform; (2) process the user profile information quantitatively, with the text in the user profile information being transformed into digital information using the ladder assignment method—the weight of the four factors in the influence index of the user profile information is calculated by the information entropy weight assignment method, and the relative reliability values are obtained by linear summing; (3) for the quantitative processing of user-generated content information, we primarily analyze the spread breadth and influence breadth of user-generated blog posts, and we calculate the relevant credibility value by weighting the features of user-generated content; (4) the two parts of the calculation results constitute the user credibility evaluation vector set, and then the vector set is divided by the soft-margin SVM algorithm to determine the user credibility evaluation results. The overall framework of the proposed method is shown in Figure 1.

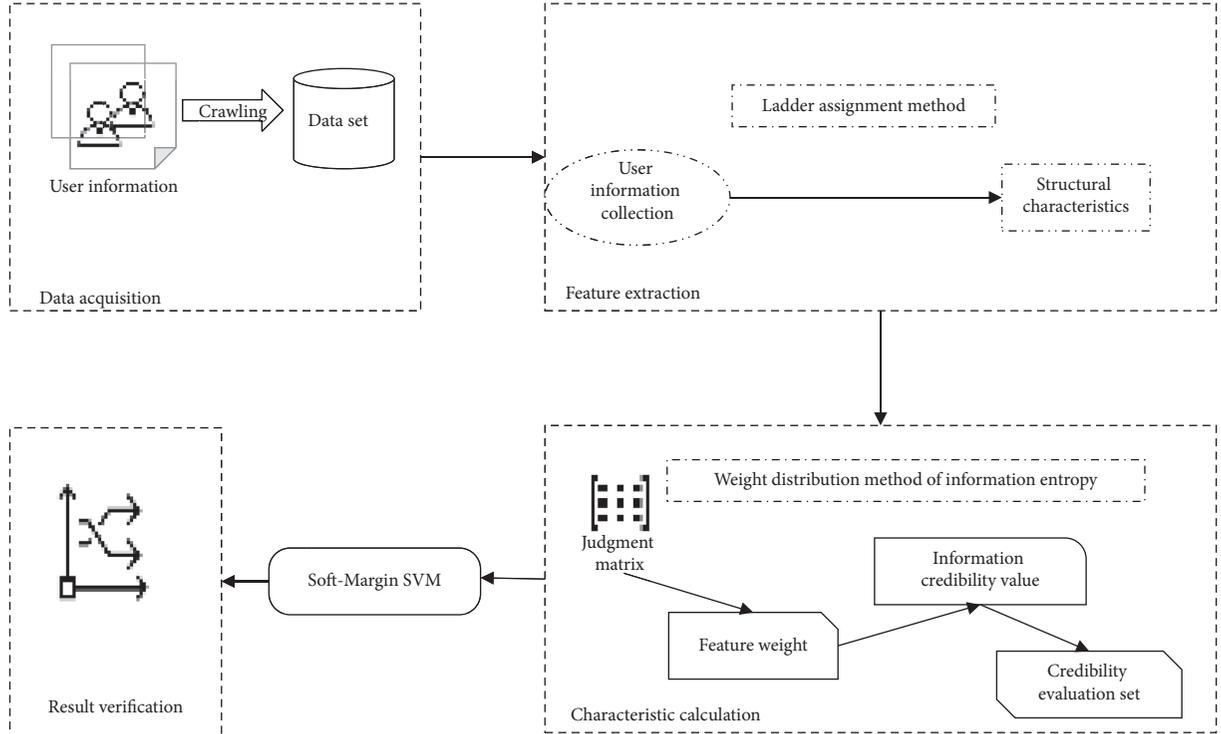


FIGURE 1: UCSSVM algorithm framework.

3.2. Quantification of User Credibility

3.2.1. Ladder Assignment of User Profile Information. The text data in the user profile information cannot be calculated mathematically. To avoid data calculation errors, while simultaneously simplifying and optimizing the calculation process, the ladder assignment of the three types of data is made according to the different characteristics of state given different values. The assignment method is shown in Table 1.

In the actual user information, the user information is divided into text type and digital type. To reduce errors after calculating the value of the two types of information, the digital-type information in the user data is assigned. The assignment methods are shown in Table 2.

3.2.2. Reliability Calculation of User Profile Information.

In social networks, user profile information reflects user authenticity and offers higher credibility than having none at all. For example, the Sina Weibo platform has a complete system of personal information. When filling out personal information, the Microblog platform has designed strict format corrections to ensure the authenticity and effectiveness of the information. The user information involved includes 20 types: 14 types of user profile information and six types of user-generated content information. User profile information includes the user nickname, UID, gender, birthday, educational background, user profile, URL, occupation, company, hometown, number of fans, number of mutual fans, number of fans, and interest tags. User-generated content information includes the number of blog posts, number of blog likes, amount of blog forwarding,

TABLE 1: Text data assignment table.

Features	State	Assignment
User nickname	System type	0
	Nonsystematic type	1
User profile	Nonexistent	0
	Exist	1
User education	Nonexistent	0
	Primary school	1
	Middle school	2
	University and above	3

TABLE 2: Numerical data assignment table.

Features	Value range	Assignment
Number of mutual fans	$0 - (M/N)$	1
	$(M/N) - (2M/N)$	2
	$(2M/N) - (3M/N)$	3
	$(3M/N) - (4M/N)$	4
	\dots	\dots
	$(N-1)(M/N) - M$	N

number of blog comments, number of blog tags, and special characters of blog posts.

The user credibility of profile information involves integrity and locality. Integrity credibility refers to the integrity of user profile information, and locality credibility refers to the influence index of user profile information. The calculation result of user profile information reliability is the linear sum of the user profile information influence index $G(i)$ and user profile integrity $UI(i)$. The reliability of user profile information $U_p(i)$ is calculated by

$$U_p(i) = UI(i) + G(i). \quad (9)$$

Definition 1. User profile information integrity. It is the ratio between the number of personal information tags that users are willing to disclose to other users in the social network and the total number of tags in the user information integrity evaluation system. It is known that the feature item of a user's integrity evaluation system is n , and the actual feature item disclosed by the user is $P_j(i)$, which is calculated as

$$UI(i) = \frac{\sum_{j=0}^n P_j(i)}{n}, \quad (10)$$

where $UI(i)$ represents the information integrity of user i as the configuration file, while n represents the total number of labels in the evaluation system for the information integrity of user i as the configuration file, and $P_j(i)$ is defined as

$$P_j(i) = \begin{cases} 1, & j = 1, \\ 0, & j \neq 1. \end{cases} \quad (11)$$

The user profile information influence index includes four characteristics: user nickname, user education, user profile, and number of mutual followers. Based on these, the authentication quad $G(F, E, P, H)$ of the user profile information influence index is constructed. The calculation method of user influence index using quaternization is as follows:

$$G(i) = n_1 F(i) + n_2 E(i) + n_3 P(i) + n_4 H(i), \quad (12)$$

where $G(i)$ represents the user influence index of user i , $F(i)$ represents the user nickname type of user i , $E(i)$ represents the education level of user i , $P(i)$ represents the profile status of user i , and $H(i)$ represents the number of mutual fans. n_1, n_2, n_3, n_4 are the weight distribution values of the user profile influence index authentication quad $G(F, E, P, H)$.

3.2.3. Credibility Calculation of User-Generated Content Information. The definition of the credibility of user-generated content $U_{\text{ucg}}(i)$ comprises two parts: the breadth of influence $IU(i)$ and the breadth of dissemination $CU(i)$ of user blog posts. The calculation equation is

$$U_{\text{ucg}}(i) = CU(i) + IU(i), \quad (13)$$

where $CU(i)$ represents the spread breadth of users' blog posts and $IU(i)$ represents the influence breadth of users' blog posts.

The user-generated content information primarily includes the text of blog posts published by users and the digital information of the interaction frequency between other users and blog content, such as the frequency information generated by the interaction behavior of forwarding, comments, and likes. Therefore, the interaction frequency data of blog posts published by target users can represent the communication breadth and influence breadth of blog posts published by users.

Definition 2. Breadth of influence of a user's blog post. It is defined as the influence degree of a user's blog on other users, which is reflected principally in the frequency of other users' liking and commenting on a target user's blog. The calculation is

$$IU(i) = \frac{1}{N} \sum_{i=1}^N \frac{D + Co}{\max(D, Co) + 1}, \quad (14)$$

where N represents the number of blog posts posted by the user, D represents the number of likes on the blog, and Co represents the number of comments on the blog. At the same time, to prevent the denominator from taking zero, 1 should be added to the value in the denominator. The larger the result of $IU(i)$ after calculation, the wider the impact of user-generated content will be.

Definition 3. The breadth of a user's blog post. The frequency of the user's blog posts being viewed by other users is measured primarily by the length of the forwarding chain of the user's blog posts—that is, the longer the forwarding chain of the user's blog posts, the wider the user-generated content is spread. The calculation is shown in:

$$CU(i) = \overline{RT} = \frac{1}{N} \sum_{j=1}^n RT(j), \quad (15)$$

where \overline{RT} represents the average forwarding chain length of the blog post published by the user, and $RT(j)$ represents the forwarding chain length of the blog post j published by the user.

3.3. Evaluation of User Credibility. For a given linearly separable data set $\{U_p(i), U_{\text{ucg}}(i)\}$, $i = 1, 2, \dots, N$. The linear discriminant function $f = \omega^T x + b$ in the two-dimensional space, the evaluation samples can be separated by the hyperplane $\omega^T x + b = 0$, where ω is the weight vector, and b is the classification threshold. To require the classification line to correctly classify all samples, it must satisfy the following: $y_i(\omega^T x + b) - 1 \geq 0$, $i = 1, 2, \dots, n$, $y_i = \pm 1$.

The hyperplane that satisfies the aforementioned conditions and maximizes the classification interval is the optimal classification surface, as shown in Figure 2:

Here, $\omega = \{\omega_1; \omega_2; \dots; \omega_d\}$ is a normal vector (which determines the direction of the hyperplane), d is the number of eigenvalues, and b is the displacement term (which determines the hyperplane and the distance between the origins). As long as the normal vector ω and the displacement b are determined, a dividing hyperplane can be determined uniquely. The distance between the hyperplane and any point on the marginal hyperplane on both sides of it is $(1/\omega)$.

To alleviate the problem of SVM overfitting and the existence of the data to be classified in the interval, we propose allowing the SVM algorithm to have reasonable errors in some classification results so that the soft-margin SVM is introduced. Specifically, the sample points in the SVM need to satisfy the algebraic equation (16) of the constraint condition.

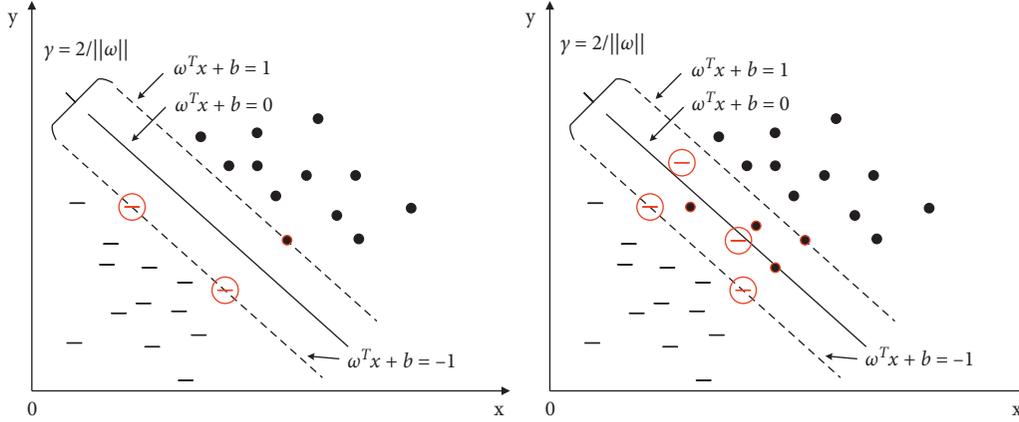


FIGURE 2: Hard-margin and soft-margin SVM segmentation hyperplane.

$$\begin{cases} \omega^T x_i + b \geq 1, & y_i = 1, \\ \omega^T x_i + b \geq -1, & y_i = -1. \end{cases} \quad (16)$$

The soft-margin SVM allows some samples not to meet the constraints because linear inseparability means that some sample points cannot meet the condition that the function interval is greater than or equal to 1, that is, $\exists i, 1 - y_i(\omega^T x_i + b) > 0$. The solution is to introduce a slack variable $\zeta_i > 0$ for each sample point. For those sample points that do not meet the constraint condition, the function interval as well as the slack variable must be made greater than or equal to 1, so our constraint condition becomes:

$$f = \min_{\omega, b} \frac{1}{2} \omega^2 + C \sum_{i=1}^m \zeta_{0/1}(y_i(\omega^T x_i + b) - 1), \quad (17)$$

where $\zeta_{0/1}$ represents the 0/1 loss function:

$$\zeta_i(z) = \begin{cases} 1, & \text{if } z < 0; \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

To optimize the soft-margin SVM and improve evaluation accuracy, slack variables are introduced into the constraint conditions and the balance coefficient C is added to the objective function to solve this problem. However, it is necessary to satisfy the function hyperplane with a maximization interval so that the samples that do not meet the constraints should be as few as possible. The balance coefficient C is also added to reconcile $(1/2)\omega^2$ and $y_i(\omega^T x_i + b) + \zeta_i \geq 1$ two-part coefficient. Therefore, the representation function of the soft-margin SVM can be written as:

$$\begin{aligned} f(\omega, b) &= \min_{\omega, b} \frac{1}{2} \omega^2 + C \sum_{i=1}^m (y_i(\omega^T x_i + b) - 1) \\ \text{s.t. } & y_i(\omega^T x_i + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, i = 1, 2, \dots, m, \end{aligned} \quad (19)$$

where $C > 0$ is called the balance coefficient. When the value of C is large, the constraint on misclassification increases, and when the value of C is small, the constraint on misclassification decreases. Among them, the value of the balance coefficient C is $C = 10^n$ and $n = -3, -2, -1, 0, 1, 2, 3$.

4. Experimental Analysis

To verify the rationality and effectiveness of the method in this paper, a comparative experiment with three other algorithms is set up. The three algorithms are derived from the literature [17, 19, 21] related to user credibility research at home and abroad. The experimental verification is based on Python and LIBSVM-related simulation platforms. The user data set used in the experiment selects user data from the social network Sina Weibo, which includes about 3,500 users with user-related profile information and generated content information. The number of experimental groups between the training set and the test set is set to 3:1, and the experimental data is grouped by the cross method.

4.1. Measurement Indicators. To evaluate the effectiveness of the user credibility calculation model and compare the advantages and disadvantages of different algorithms for measuring user credibility, the accuracy rate, precision rate, recall rate, and balanced $F1$ score are introduced as evaluation indicators. At the same time, the accuracy decline rate and average accuracy rate are introduced to characterize the robustness of user credibility evaluation algorithm.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (20)$$

Accuracy means correctly predicting the proportion of the number of (TPs) and (TNs) to the total number of experimental users.

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (21)$$

Precision indicates the probability of correctly marking a TP among the samples marked as TPs in the marking results.

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (22)$$

The recall rate represents the probability that a TP in the original data set will be correctly marked as a TP at the end.

$$F_1 = \frac{2 \text{ Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (23)$$

The balance $F1$ score is the harmonic average of the precision rate and recall rate.

$$\text{Decline rate} = \frac{\max_{\text{accuracy}} - \min_{\text{accuracy}}}{N_{\max} - N_{\min}}. \quad (24)$$

The accuracy decline rate is the ratio of the difference in accuracy to the difference in the number of samples.

$$\text{Accuracy average} = \frac{\sum_{i=1}^n \text{accuracy}_n}{n}. \quad (25)$$

The accuracy average is the average value of the accuracy of user reliability evaluation.

Among them, the number of samples $N = \text{trusted users} + \text{malicious users} = P + N = \text{TP} + \text{FP} + \text{FN} + \text{TN}$, where TP means “real,” and TPs are marked as TPs, FP means “real” and TNs are marked as TPs, TN means that a real TNs is marked as a TPs. $\max_{\text{accuracy}} - \min_{\text{accuracy}}$ is the difference in accuracy, $N_{\max} - N_{\min}$ is the difference in the number of samples.

4.2. Analysis of Experimental Results

4.2.1. The Impact of the Balance Coefficient C on User Credibility Evaluation. The balance coefficient C refers to the weight between controlling the hyperplane with the largest interval of the SVM and ensuring the smallest deviation of the data points. It can be inferred from the principle of a soft-margin SVM that $C > 0$, C is a coefficient that balances $(1/2)\omega^2$ and the error interval, so $C = 10^n$, $n = -3, -2, -1, 0, 1, 2, 3$. The best value of the balance coefficient C is obtained through fivefold cross-validation of experimental groups, as shown in Table 3.

The balance coefficient C is selected, and the balance coefficient C with the highest accuracy is obtained based on the fivefold cross-validation. The larger the value of the balance coefficient C , the smaller the value of the required slack variable, and the smaller the evaluation function hyperplane interval, that is, the smaller the tolerance to noise, the higher the accuracy of the evaluation result. The smaller the value of the balance coefficient C , the larger the value of the required slack variable, and the larger the evaluation function hyperplane interval, that is, the greater the tolerance to noise, the smaller the accuracy of the evaluation result.

It can be seen from Table 4 that on the one hand, to ensure the accuracy of user credibility evaluation, a larger value of the balance coefficient C is required; on the other hand, a certain degree of noise tolerance is required to avoid overfitting of the evaluation function, and a balance coefficient C as small as possible is required. Comprehensive analysis and experimental results show that when the balance coefficient is 10.00, the accuracy of user credibility evaluation is the highest, and the balance coefficient C is the smallest, as shown in Table 4.

TABLE 3: The best value of the balance coefficient C .

Balance coefficient C	Accuracy					Average value
	1	2	3	4	5	
0.001	0.900	0.900	0.900	0.925	0.900	0.905
0.010	0.900	0.875	0.900	0.925	0.925	0.905
0.100	0.875	0.900	0.900	0.925	0.900	0.900
1.000	0.900	0.900	0.900	0.925	0.925	0.910
10.00	0.925	0.900	0.900	0.925	0.950	0.920
100.0	0.925	0.900	0.900	0.925	0.950	0.920
1000	0.925	0.900	0.900	0.925	0.950	0.920

TABLE 4: Average value of the best value of the balance coefficient C .

C	0.001	0.01	0.1	1	10	100	1000
Average value	0.905	0.905	0.900	0.910	0.920	0.920	0.920

4.2.2. The Impact of the Proposed Method on User Credibility Evaluation. To compare the effectiveness and rationality of the proposed algorithm for user credibility evaluation, this paper selects three user credibility evaluation methods and the method UCSVM proposed in this paper to evaluate user credibility and compare their results.

One paper [17] adopts the decision tree algorithm in machine learning to learn the user profile information quantitatively to achieve the purpose of user reliability evaluation. Another paper [19] characterizes user credibility by quantifying user-generated content information. The final paper [21] comprehensively considers multiple types of user information and uses a PageRank algorithm to quantify user information, thereby evaluating user credibility. The effectiveness of UCSSVM is illustrated from four aspects, such as Figure 3(a) accuracy, Figure 3(b) precision, Figure 3(c) recall rate, and Figure 3(d) balanced $F1$ score in Figure 3. In the papers [19, 21], user-generated content information contains more information used in user credibility evaluation. Specific information can characterize user credibility more accurately, so the user credibility evaluation result of paper [21] is better than that of paper [19]. Paper [21] uses a PageRank algorithm to quantify user information, which comprehensively considers various types of user information to represent user credibility and avoids the problem of insufficient accuracy of user credibility evaluation results caused by sparse user information in papers [19, 21]. Because papers [17], [19], and [21] evaluate user credibility in the linear summation dimension, there is a problem of user aliasing at the user classification threshold. It can be seen from Figure 3 that the user credibility evaluation method UCSSVM proposed in this paper is superior to the other three algorithms in relation to the user credibility evaluation results because our proposed method can allocate the corresponding weight reasonably according to the importance of user data. At the same time, the user credibility evaluation in the two-dimensional plane avoids the problem that the linear summation risks causing in relation to the aliasing of TPs and TNs at the classification threshold. Because there is a negative correlation between the number of users and the evaluation index, that is, when the number of users continues to increase, the number of users in the interval

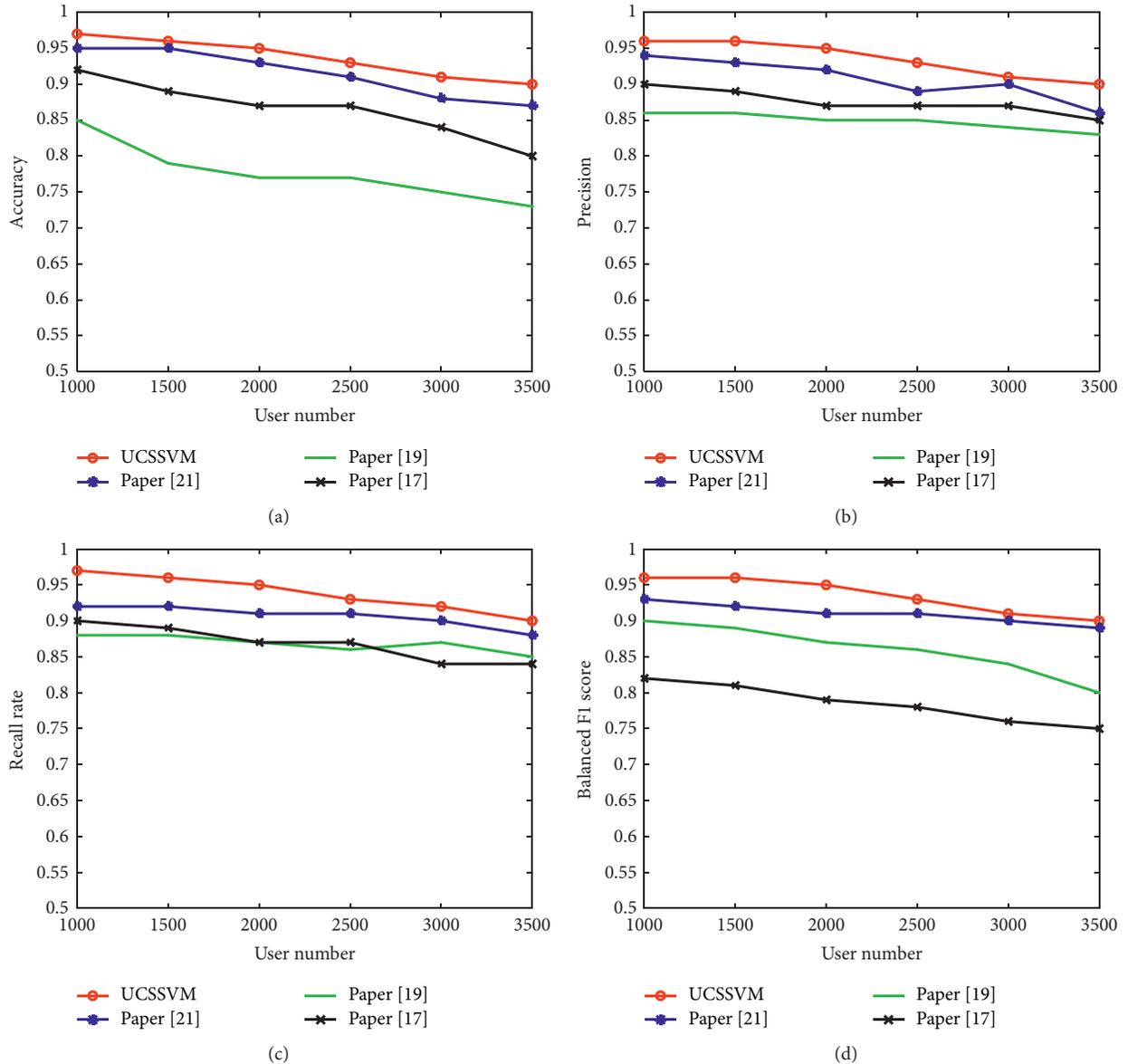


FIGURE 3: Evaluation results of UCSSVM and the three methods. (a) The accuracy of the evaluation result. (b) The precision of the evaluation result. (c) Recall rate of evaluation results. (d) Balanced $F1$ score of the evaluation result.

hyperplane of UCSSVM increases—the relaxation variable is made larger and the tolerance to the noise data smaller. This, in turn, makes the error of the user evaluation result larger, resulting in a downward trend of the evaluation index.

As can be seen from Tables 5 and 6, the user credibility evaluation method UCSSVM proposed in this paper is better than the other three algorithms in the decline rate and average value of the accuracy of user credibility evaluation results. The increase of the number of experimental samples will increase the positive proportion of the amount of user information and the divergence of user information, which leads to the decline of the evaluation accuracy of the user credibility evaluation algorithm, but the method proposed in this paper integrates multiple types of user information and weakens the impact of the increase of the amount of user information and the divergence of user information on the

TABLE 5: Decline rate of accuracy.

	UCSSVM	Paper [21]	Paper [19]	Paper [17]
Decline rate	0.07	0.08	0.13	0.14

The decline rates of the accuracy of user credibility evaluation results for papers [17], [19], [21], and UCSSVM, are 0.13, 0.14, 0.08, and 0.07 respectively, the average values of the evaluation results are 0.875, 0.848, 0.907, and 0.935 respectively.

accuracy of user credibility evaluation results. The distribution of user feature weight based on information entropy reduces the impact of the change of information on the accuracy of user credibility evaluation so as to ensure the stability of the algorithm. Therefore, compared with the other three algorithms, UCSSVM shows the lowest decline rate of accuracy and the highest average value of accuracy. This method has better robustness.

TABLE 6: Average value of accuracy.

	Accuracy						Average value
	1000	1500	2000	2500	3000	3500	
UCSSVM	0.96	0.96	0.95	0.93	0.91	0.90	0.935
Paper [21]	0.94	0.93	0.92	0.89	0.90	0.86	0.907
Paper [19]	0.86	0.86	0.85	0.85	0.85	0.83	0.85
Paper [17]	0.90	0.89	0.87	0.87	0.87	0.85	0.875

5. Conclusion

Aiming at the problem of aliasing of user credibility results in the linear summation dimension, a user evaluation method UCSSVM based on a soft-margin SVM is proposed in this paper. The judgment matrix in the information entropy weight distribution method is used to solve the weight distribution problem of four feature items in the credibility calculation of user-generated content information. The user text and digital information are processed by the ladder assignment method, which reduces the error of user credibility calculation results caused by different types of information. Finally, the set of user profile information credibility and user-generated content information credibility is evaluated. The experimental results show that the user credibility evaluation method UCSSVM proposed in this paper not only avoids the user aliasing easily caused by the linear evaluation algorithm but it also improves the accuracy of the user evaluation results.

In the future research, we can further improve the existing work and in-depth research through the following points: (1) we can explore the optimal application of more machine learning algorithms in the selection and quantification of user information related features. (2) Explore the optimization of evaluation results and the reduction of related computing overhead by different support vector machine variant algorithms, explore more user credibility evaluation models, and improve the robustness and accuracy of the evaluation algorithm. (3) The extraction of user information features in user credibility evaluation is deeply studied to optimize the best balance between the amount of feature information and the number of features, and simplify the computational complexity of user credibility.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by Program for Innovative Research Team in University of Henan Province (21IRTSTHN015).

References

- [1] M. Alkhamees, S. Alsaleem, M. Al-Qurishi, M. Al-Rubaian, and A. Hussain, "User trustworthiness in online social networks: a systematic review," *Applied Soft Computing*, vol. 103, no. 2, Article ID 107159, 2021.
- [2] D. Varshney and D. K. Vishwakarma, "A review on rumour prediction and veracity assessment in online social network," *Expert Systems with Applications*, vol. 168, no. 4, Article ID 114208, 2020.
- [3] C. Rashmi and M. M. Kodabagi, "Profiling of social network users using proximity measures," in *Proceedings of the 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, pp. 24–28, Bengaluru, India, October 2020.
- [4] M. Alrubaian, M. Al-Qurishi, A. Alamri, M. Al-Rakhami, M. M. Hassan, and G. Fortino, "Credibility in online social networks: a survey," *IEEE Access*, vol. 7, pp. 2828–2855, 2019.
- [5] X. Yi, E. Bertino, F.-Y. Rao, K.-Y. Lam, S. Nepal, and A. Bouguettaya, "Privacy-preserving user profile matching in social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1572–1585, 2020.
- [6] Y. Xia, L. Pan, L. Shi, and F. Zou, "Attribute credibility based sybil group detection in online social networks," in *Proceedings of the 2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, pp. 358–363, Changsha, China, June 2016.
- [7] T. Khan and A. Michalas, "Trust and believe—should we? evaluating the trustworthiness of twitter users," in *Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 1791–1800, Guangzhou, China, January 2021.
- [8] A. N. Ramkissoon and S. Mohammed, "An experimental evaluation of data classification models for credibility based fake news detection," in *Proceedings of the 2020 International Conference on Data Mining Workshops (ICDMW)*, pp. 93–100, Sorrento, Italy, November 2020.
- [9] F. Amato, V. Moscato, A. Picariello, and G. Sperli, "Recommendation in social media networks," in *Proceedings of the 2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pp. 213–216, Laguna Hills, CA, USA, April 2017.
- [10] F. Zhou, J. Jin, X. Du, B. Zhang, and X. Yin, "A calculation method for social network user credibility," in *Proceedings of the 2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, Paris, France, May 2017.
- [11] M. Al-Qurishi, M. S. Hossain, M. Alrubaian, S. M. M. Rahman, and A. Alamri, "Leveraging analysis of user behavior to identify malicious activities in large-scale social networks," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 799–813, 2018.
- [12] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurrency and Computation: Practice and Experience*, vol. e5909, 2020.
- [13] Y. Cardinale, I. Dongo, G. Robayo, D. Cabeza, A. Aguilera, and S. Medina, "T-CREo: a twitter credibility analysis framework," *IEEE Access*, vol. 9, pp. 32498–32516, 2021.
- [14] M. Viviani and G. Pasi, "Credibility in social media: opinions, news, and health information—a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 5, 2017.

- [15] J. Yang, M. Yu, Q. Han, M. Lu, and C. Yang, "A twitter data credibility framework—hurricane harvey as a use case," *International Journal of Geo-Information*, vol. 8, no. 3, 2019.
- [16] D. Hassan, "A text mining approach for evaluating event credibility on twitter," in *Proceedings of the 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pp. 171–174, Paris, France, June 2018.
- [17] A. Narayanan, A. Garg, I. Arora, T. Sureka, M. Sridhar, and H. B. Prasad, "IronSense: towards the identification of fake user-profiles on twitter using machine learning," in *Proceedings of the 2018 Fourteenth International Conference on Information Processing (ICINPRO)*, pp. 1–7, Bangalore, India, December 2018.
- [18] E. Van Der Walt, J. Eloff, and J. Eloff, "Using machine learning to detect fake identities: bots vs humans," *IEEE Access*, vol. 6, pp. 6540–6549, 2018.
- [19] H. Slimi, I. Bounhas, and Y. Slimani, "URL-based tweet credibility evaluation," in *Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–6, Abu Dhabi, UAE, November 2019.
- [20] V. P. K. Verma, V. Sharma, and S. Agarwal, "Credibility investigation for tweets and its users," in *Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 925–928, Erode, India, March 2019.
- [21] Q. Sun, N. Wang, Y. Zhou, and Z. Luo, "Identification of influential online social network users based on multi-features," *International Journal of Pattern Recognition & Artificial Intelligence*, vol. 30, no. 6, Article ID 1659015, 2016.
- [22] S. Zhang, Y. Cai, and H. Xia, "A privacy-preserving interactive messaging scheme based on users credibility over online social networks," in *Proceedings of the 2017 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1–6, Qingdao, China, October 2017.
- [23] P. Li, W. Zhao, J. Yang, Q. Z. Sheng, and J. Wu, "Let's CoRank: trust of users and tweets on social networks," *World Wide Web*, vol. 5, no. 23, pp. 2877–2901, 2020.
- [24] W. Deng, S. Shang, X. Cai, H. Zhao, Y. Song, and J. Xu, "An improved differential evolution algorithm and its application in optimization problem," *Soft Computing*, vol. 25, no. 7, pp. 5277–5298, 2021.
- [25] R. Rahangdale and U. Thakar, "A user action based approach to determine trustworthiness among users in social network," in *Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 148–152, Palladam, India, February 2017.
- [26] J. Khan and S. Lee, "Implicit user trust modeling based on user attributes and behavior in online social networks," *IEEE Access*, vol. 7, pp. 142826–142842, 2019.
- [27] Q. Zheng and S. Qu, "Credibility assessment of mobile social networking users based on relationship and information interactions: evidence from China," *IEEE Access*, vol. 8, pp. 99519–99527, 2020.
- [28] N. A. Utami, W. Maharani, and I. Atastina, "Personality classification of facebook users according to big five personality using SVM (support vector machine) method," *Procedia Computer Science*, vol. 179, no. 1, pp. 177–184, 2021.
- [29] B. B. Hazarika and D. Gupta, "Density-weighted support vector machines for binary class imbalance learning," *Neural Computing and Applications*, vol. 33, no. 9, pp. 4243–4261, 2020.