A Machine Learning Based Security Detection Method for Privacy Data in Social Networks

Zhiyu Huang(✉) and Chenyang Li

Shenyang Institute of Technology, Shenyang 113122, China
huangzhiyu2004@163.com

Abstract. In order to improve the social Internet privacy data security detection effect and improve the data security detection efficiency, this paper proposes a social Internet privacy data security detection method based on machine learning. First, collect social Internet privacy data and construct N-Gram language model to realize the standardization of social Internet privacy data; Secondly, a semantic vector based representation model is used to obtain topic semantic vectors, and the obtained topic semantic vectors are matched; Finally, social Internet privacy data security risk detection is carried out by using the skew coefficient method in machine learning. The results show that the method in this paper effectively compresses the time consumption of security detection through machine learning. The time consumption of detection is only 5.3 s, and the accuracy of data detection can reach 99.5%. The method in this paper can effectively improve the efficiency of social Internet privacy data security detection and improve the detection accuracy, but the detection cost needs to be reduced.

Keywords: N-Gram language model · Semantic vector · Deviation coefficient · machine learning

1 Introduction

In recent years, with the rapid development of social network, in view of the social network data publishing technology applications in some studies and has made great progress in the social network to provide a large number of social network user data, the data is collected, grind Gui and release [1] for a variety of purposes. Some large Internet companies can combine large amounts of data to construct a clear behavioral map of an individual, and then predict their preferences and behaviors [2]. This data is very valuable in the consumer market, and can be used to proactively promote certain products or services to specific groups of people. However, a lot of user data in social networks also contains many users' personal privacy information, such as personal address information, friends, interests and hobbies, etc. In addition, it is difficult for network service providers to restrict the use of data at present. Because the user's identity, address, contact information and other information are currently in the hands of service providers, but how the people who deal with these data every day should use the

L. Yun et al. (Eds.): ADHIP 2023, LNICST 547, pp. 350–365, 2024.
https://doi.org/10.1007/978-3-031-50543-0_24

data to avoid privacy disclosure is a very worthy question. Once these released data is maliciously used, it will cause unpredictable loss and impact on individuals and society. Therefore, how to obtain personal privacy protection while releasing social network user data is a hot issue in current research.

In recent years, due to the rapid popularity of social network services, the amount of social network data has exploded in various application scenarios. Since all data is generated from people's daily behavior, almost all data is related to personal data. If these data are correctly shared or aggregated, they can be used to extract valuable information and knowledge. However, while sharing or mining data, there may be privacy breaches due to the failure to control personal data from users. Therefore, data mining for social network privacy protection has received great attention in recent research. Reference [3] proposed a social Internet privacy privacy data security detection method based on cloud computing, which uses encryption algorithms to encrypt users' privacy data to ensure the security of data during transmission and storage. Use access control policies to restrict access to private data. Using methods such as role-based access control (RBAC) or attribute based access control (ABAC) to manage and control user access to data. Classify and label private data based on sensitivity and privacy level, and set different access permissions and security policies for different data. Record user access and operations to private data, and generate audit logs. By monitoring and analyzing audit logs, abnormal behavior and security vulnerabilities can be detected in a timely manner. Regularly backup private data and establish a comprehensive data recovery mechanism to prevent data loss or damage. Establish a real-time security monitoring system to monitor and analyze data traffic and user behavior on social network platforms. Once abnormal activities or potential security threats are detected, timely alerts are issued and corresponding response measures are taken. Using network security devices such as firewalls and intrusion detection systems (IDS) to strengthen network security protection for social network platforms and prevent unauthorized access and attacks. Regularly conduct security assessment and Penetration test, find potential security vulnerabilities and risks, and timely repair and strengthen security measures. This method can effectively improve the accuracy of social Internet privacy privacy data security detection, but the efficiency of privacy data security detection is low. Reference [4] proposes a social network user data security protection method based on the Big data background, classifies and marks social network user data, classifies them according to sensitivity and privacy level, and sets different protection policies for different data. Automated methods can be used to identify and label sensitive information in user data. Desensitize some non sensitive attributes to protect users' personal privacy. Data anonymization user data to remove directly identifiable personal identity information to protect user privacy. The Differential privacy technology is used to realize the de identification of data and protect the personal privacy of users. Establish a comprehensive access control mechanism to restrict access to user data. Authorize different users and roles through the role and permission management system, ensuring that only authorized users can access the corresponding data. Establish a real-time security monitoring system to monitor and analyze data traffic and user behavior on social network platforms. Discovering abnormal behavior and potential security threats through the use of machine learning and data mining techniques. Regularly backup user data, establish a comprehensive

data recovery mechanism, adopt distributed storage and disaster recovery technology, and improve data reliability and availability. However, this method has poor security detection efficiency.

Currently, many social network service providers have realized the importance of the personal data they collect for making business decisions, information discovery, or other research needs. However, these data typically contain private and sensitive information about individuals, which can be discovered through privacy breach attacks on published data [5]. Therefore, these data should be published in a manner that prohibits the disclosure of any personal privacy information and the disclosure of personal identity. In addition, social network service providers should publish data to the public without violating personal information protection guidelines. This has raised concerns about protecting personal privacy when publishing data. Therefore, the publication of privacy protection data on social networks At present, the issue of privacy protection for data release on social networks is mainly how to release data in a way that protects personal privacy. With the deepening of the research, more valuable research results have been obtained at home and abroad. The research results can be divided into two categories: one is to achieve the purpose of privacy protection based on K-anonymity model. A common method is to use non-specific information to replace sensitive and specific information, that is, information generalization [6]. The other is to use probability or statistical methods to protect data privacy while keeping the statistical characteristics and classification attributes of the final data unchanged. Such as clustering, randomization, sampling, data exchange, and data perturbation for data publication.

To solve the problem of poor security detection efficiency and poor detection accuracy, this paper proposes a social Internet privacy data security detection method based on machine learning, constructs N-Gram language model to realize social Internet privacy data standardization, uses semantic vector based representation model to obtain topic semantic vector, and matches the obtained topic semantic vector, Social Internet privacy data security risk detection based on the method of partial coefficient in machine learning.

2 Machine Learning Based Privacy Data Security Detection Method for Social Networks

2.1 Access to Social Network Privacy Data

The data set used in this paper comes from public data sets and web crawler. There are more than 4 million data sets collected, of which about 3.8 million and 200000 are public and crawlers respectively. The log data collected on social networking sites generally includes user ID, number of followers, number of followers, number of followers, timestamp, and message text. The public experimental data used in this article comes from Baidu AIStudio, which is an artificial intelligence learning and training community based on the Baidu deep learning platform Feijiang, providing a variety of massive data for users to use [7]. All types of data are freely uploaded by professionals in various fields, and the platform will conduct review and supervision, with certain standardization of the data. The data we downloaded on this platform is taken from Weibo. Weibo is the

largest social media platform in China, with over 550 million active users in 2020. Users can register Weibo accounts, create personal information, establish friendly contacts with other users, upload photos, share interesting things, etc. The platform has a certain degree of user stickiness and strong research value.

After a user registers an account on Weibo platform, the system will assign a unique ID number to the user as the user ID, and the user ID needs to be passed when the platform uses the open API to obtain the corresponding data. Therefore, we obtain user data by ID, but the user ID of the platform is not continuous. In order to reduce sampling deviation, We start by randomly generating 10,000 ids. The API was used to obtain the personal information of these ids. Among them, 1,547 ids returned correct information, indicating that these ids did not correspond to users. For the accounts that returned the correct information, we used the API to get the dynamic content published in the last week.

Next, it is necessary to determine whether obtaining account information is abnormal. For this purpose, two classmates were invited to do manual marking.

These two volunteers are both graduate students in our laboratory and familiar with the dynamic content of Weibo. Volunteers determine whether it is abnormal behavior based on the content posted by the user. If an account repeatedly forwards, posts useless information, or points to websites such as shopping, phishing, or pornography in a short period of time, then the main body of the account can be determined to be abnormal based on the abnormal behavior.

Volunteers need to manually identify all IDs, considering that manual identification may result in errors, and therefore abandon controversial accounts. In the end, 1045 accounts were identified as abnormal, and the 120780 message text associated with them was also marked as abnormal. The remaining 8067 IDs were identified as normal accounts, and the 16450 message content associated with them was identified as normal.

2.2 Standardized Processing of Social Network Privacy Data Based on N-Gram Language Model

The original data on the social network platform exists in the form of text, and the content of each message in the serialized data is text. In the previous section, the linguistic probability model of the (n-1) Markov chain in the collected messages is presented according to the probability judgment statement of the occurrence of n word. Its basic idea is to divide the text content into sliding Windows of length N [8] according to byte size. Each byte fragment is called a *gram*, and the occurrence frequency of all *gram* is counted, and the threshold is set in advance for filtering, so as to form a list of *gram*, that is, the feature space of text vectors. Each *gram* in the result is a feature vector dimension.

This model assumes that the n word is only related to the occurrence of the first $n-1$ words, and is independent of other words. The product of the probability of each word's occurrence is used as the probability of the entire sentence. The commonly used models include the binary model Bi Gram and the ternary model Tri Gram. The specific usage details will be introduced below.

If there is a sequence composed of m words, hoping to obtain the probability $p(w_1, w_2, ..., w_i)$, according to the chain rule, we can obtain

$$p(w_1, w_2, ..., w_m) = p(w_1) * p(w_2|w_1) * p(w_3|w_1, w_2)...p(w_m|w_1, w_2, ..., w_{m-1}) \quad (1)$$

Among them, w_i represents the probability of a certain word appearing. This paper uses the Markov hypothesis to calculate this probability, taking into account that the current word is only related to a limited number of words before it, so there is no need to go back to the first word in the calculation, which can greatly reduce the calculation of the above formula. After simplification, the above formula can be expressed as:

$$p(w_1, w_2, ..., w_m) = p(w_i|w_{i-n+1}, ..., w_{i-1}) \tag{2}$$

When n = 1, the unigram model can be represented as:

$$p(w_1, w_2, ..., w_m) = \prod_{i=1}^{m} p(w_i) \tag{3}$$

When n = 2, bigram model can be expressed as:

$$p(w_1, w_2, ..., w_m) = \prod_{i=1}^{m} p(w_i|w_{i-1}) \tag{4}$$

With the above model concept, the above conditional probability value can be calculated from the given training corpus by using Bayes theorem.

For the bigram model:

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{5}$$

For the N-Gram model:

$$p(w_i|w_{i-n-1}, ..., w_{i-1}) = \frac{C(w_{i-n-1}w_i)}{C(w_{i-n-1}w_{i-1})} \tag{6}$$

In this paper, 2-g model is used. The specific modeling method is designed to define a sliding window with length of 2 and move back step by step. After each move, 2 words in the window are taken as a sequence, and finally the sequence is converted into a numerical vector. Since the size of the sliding window is set to 2 and there are a maximum of 256 characters in the text, there will be 2**16 2-g combinations [9]. A common practice is to use a one-hot vector (set to 1 if present, 0 otherwise). To preserve more information, it is also possible to replace 1 and 0 with the number of times each 2-g appears in the text. Now, each message content can be transformed into a vector. Since the content of each text is different, the length of each text is naturally inconsistent. To ensure that the input data has the same length, you can fill it with specific characters to ensure that the input content length is standardized.

2.3 Semantic Vector Based Representation Model

The original text is composed of a series of symbols, which is a set of symbols with certain semantics. The text is recorded by human beings. Due to the diversity of human language, it is quite difficult for machines to understand the text, and many obstacles

need to be overcome to accurately understand natural language. Since machines cannot understand natural language like human beings, they need to model text semantics and realize semantic analysis and understanding through solving mathematical models [10]. The model based on semantic vector maps the text vector processed above to the potential semantic space, and represents the semantic relationship through the new text vector.

For the construction of the original text vector, first extract the subject words from the preprocessed message content text vector. In this paper, we use Latent Dirichlet Allocation (LDA) topic model to extract topic words, and construct D * Z dimension text topic matrix and Z * W dimension topic word matrix according to the distribution of vocabulary content.

All parameters in PLSA model represent random variables, without considering prior distribution. Since the parameters in PLSA correspond to multinomial distribution, naturally a better choice is Drichlet distribution. We can add prior distribution before the parameters to transform PLSA model into LDA model, so that the model has strong generalization ability. At the same time, the prior distribution is added to make LDA model have a complete Bayesian framework. $\vec{\alpha}$ is a K-dimensional vector, used to represent the hyperparameter of Document-Topic distribution, and $\vec{\theta_m}$ is an K-dimensional vector composed of $p(z|d = i)$. $\vec{\beta}$ is a K-dimensional vector used to represent the hyperparameter of the Topic-Word distribution, and $\vec{\theta_k}$ is an K-dimensional vector composed of $p(w|z = i)$.

Similar to the Unigram Model, an V_i-sided die, with one word for each side, is used to represent the i document. Therefore, the probability model diagram can be decomposed into two processes, as shown in Table 1.

Table 1. LDA model generation process

LDA model generation process	
1	$\alpha \rightarrow \theta_m \rightarrow z_{m,n}$ This process means that the topic distribution θ_m of the document in Chapter m is generated by sampling from the Dirichlet distribution α, and then the topic $z_{m,n}$ of the nth word of the document in Chapter m is sampled from the topic polynomial distribution θ_m
2	$\beta \rightarrow \varphi_k \rightarrow w_{m,n}\|k = z_{m,n}$. This process means that the word distribution φ_k with the subject $z_{m,n}$ is generated by sampling from the Dirichlet distribution, and then the word $w_{m,n}$ is generated by sampling from the polynomial distribution D

In the whole model, the joint distribution of all implicit variables and visible variables is:

$$p(w_m, z_m, \theta_m, \phi|\alpha, \beta) = \prod_{n=1}^{N_m} p(\theta_m|\alpha)p(\phi|\beta)p(w_{m,n}|\varphi z_{m,n}) \tag{7}$$

In the above equation, all documents of $\Phi = \{\Phi_k\}_{k=1}^{K}$ are shared, and the maximum likelihood estimation of the word distribution in the final document is summed by z_m and the above equations θ_m and Φ are integrated:

$$p(w_m|\alpha, \beta) = \int_{\theta_m} \int_{\phi} \sum_{z_m} p(w_m, z_m, \theta_m, \phi|\alpha, \beta) \tag{8}$$

With the joint probability distribution, Gibbs Sampling algorithm can be used for sampling. First, all the words in the document are traversed and assigned A topic, i.e., $z\_(m, n) = k \sim Mult(1/K)$, where K represents the total number of topics and $z_{m,n}$ represents the n topic in the document of article m. After that, corresponding n_m^k, n_m, n_k^k, n_k represents the number of occurrences of theme k in document m, the number of topics in document m, the number of words t corresponding to topic k, and the total number of topic k. The most important Gibson sampling formula is obtained according to the existing joint probability distribution$p(\vec{w}, \vec{z})$

$$p(z_i = k|z_{\neg i}, w) \propto \frac{(n_{k,\neg i}^{(t)} + \beta_{\mathrm{t}})(n_{m,\neg i}^{(k)} + \alpha_{\mathrm{k}})}{(\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_{\mathrm{t}})} \tag{9}$$

After multiple iterations, the output Topic Word parameter matrix $\varphi_{k,t}$ and Document Topic matrix $\theta_{n,k}$ converge, as shown in Eq. (10). During the inference stage, keeping $\varphi_{k,t}$ unchanged and updating $\theta_{n,k}$ can obtain potential topic semantic vectors corresponding to the new text.

$$\varphi_{k,t} = \frac{(n_k^{(t)} + \beta_t)}{(n_k + \beta_t)} \tag{10}$$

$$\theta_{m,k} = \frac{(n_m^{(t)} + \alpha_k)}{(n_m + \alpha_k)} \tag{11}$$

The topic semantic vector is obtained, and the topic semantic vector is matched.

2.4 Semantic Model Based on Text Similarity

After the above operations, the theme words in each original text can be obtained. Furthermore, the theme words in the Document Topic matrix and Word Topic matrix can be mapped with the corresponding words in the text one by one to the user's behavioral concept, and the words and theme words can be corresponding to the user's actual behavioral concept, and they can be regularized. Based on the mapping relationship constructed between concepts, the analysis of the relationship between words and themes can be transformed into analysis between conceptual levels. At this point, what is represented in the matrix is no longer a simple distribution relationship between Word, Topic, and Document, but a semantic relationship at the conceptual level. This relationship matrix better represents the relationship between log text and user behavior concepts, which can reduce useless features and dimensions.

The closer the similarity between the words in the text and the extracted topic, the better the current word reflects the meaning of the document and is therefore worth keeping. Common methods for calculating text similarity include reference method and word shift distance method. The former first calculates the average embedding value of all words in the sentence, and then calculates the cosine similarity between the two sentences to judge. The latter uses word embedding between two texts, and then calculates the minimum distance needed to move a word from one text to another text in a semantic space vector. In this paper, text2vec package is used for judgment. This library uses Tencent AI Lab open source library to collect a variety of high-quality Chinese word data, which can achieve lower granularity of words and higher accuracy, and is very suitable for Chinese similarity calculation.

Assuming $z_{m,0}$ is the theme word of document d_m in article m, $n_{m,0}$ is the number of occurrences of theme $z_{m,0}$ in text d_m, $w_i \in (w_1, w_2, ..., w_n)$ is any word in text D, and the semantic increment between word w_i and the theme word $z_{m,0}$ in article m is $SI_{m,i} = si_{m,i} * n_{l,0}$. Where $SI_{m,i}$ is the semantic similarity between the word w_i and the d_m topic $z_{m,0}$.

3 Social Network Privacy Data Security Risk Detection

3.1 Analysis of Chaotic Sequence Encryption Process

Chaos phenomenon is a kind of random process in nonlinear deterministic system. Two particularly close original values are introduced into the same chaotic function to carry out iterative calculation. After the calculation in a specific stage, the numerical sequence has no similarity. This encryption method belongs to the deterministic system, but it is difficult to predict it, hidden in the jumbled system but can not be decomposed.

The characteristics of chaotic signal, such as non-periodicity, continuous broadband spectrum and similarity to noise, make it possess natural hidden characteristics, highly sensitive to original conditions, and long-term unpredictability, making it difficult for data to be subjected to malicious damage and attack.

Logistic mapping represents nonlinear chaotic equation, and its mapping process is as follows:

$$X_{n+1} = bX_n(1 - X_n) \quad X_n \in [0, 1] \tag{12}$$

In the equation, b represents the control parameter variable. After specifying the specific value of b, a clear time series $X_1, X_2, \cdots, X_n$ can be iteratively calculated using the random original value $X_0 \in [0, 1]$.

The security key of chaotic sequence encryption depends on chaotic key stream. In chaotic encryption system, random sequence $\{x_i\}$ generated by chaotic system is regarded as bitwise operation of key stream $\{k_i\}$ and plaintext data stream $\{m_i\}$, and then ciphertext data stream $\{c_i\}$ is obtained. The plaintext data stream is binary, while the key stream $\{k_i\}$ is obtained by data processing of the chaotic sequence $\{x_i\}$. The initial chaotic data $\{x_i\}$ is processed by computer technology, and the randomness of chaotic sequence is completed under the premise of limited computational accuracy.

Chaos encryption cipher as a sequence cipher. The encryption and decryption ends of the chaotic sequence cipher system are two completely independent and equal chaotic systems, and there is no coupling relationship between the two systems. Clear text data is encrypted at the encryption end and directly transmitted to the decryption end. The decryption end can perform decryption after receiving all data. The encryption method of chaotic sequence cipher is relatively flexible, which can effectively utilize the characteristics of chaotic signals to obtain complex encryption functions. The chaotic sequence cipher system is shown in Fig. 1.

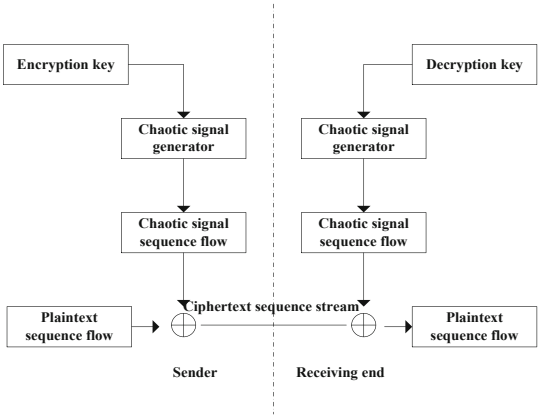

Fig. 1. Chaotic encryption process

3.2 Privacy Encryption of Social Network Privacy Data in Chaotic Sequence

Transforming the original data through chaotic sequences allows unauthorized users to obtain encrypted data, but due to the inability to decrypt it, the method for determining the information content is still unclear. Maximize data security and prevent data leakage. Simultaneously deprivarizing data before encryption can more effectively improve data security.

Record the sequence probability distribution function generated by the logistic mapping of Eq. (9) pattern as:

$$\rho_{(x)} = \begin{matrix} \pi \\ 0 \end{matrix} \begin{matrix} \dfrac{1}{x(1-x)} & 0 < x < 1 \\ else \end{matrix} \tag{13}$$

Some important statistical features of the chaotic sequence generated by Logistic mapping can be easily obtained by using $\rho_{(x)}$. For example, the average time of x, which is also the mean value of the trajectory points of the chaotic sequence, is described as:

$$\bar{x} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N} x_i = \int_0^1 x\rho_{(x)} dx = 0.5 \tag{14}$$

For the cross correlation function, if two original values x_0 and y_0 are selected separately, the sequence cross correlation function is:

$$\begin{aligned} c(l) &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} (x_i - \overline{x})(y_{i+1} - \overline{y}) \\ &= \int_0^1 \int_0^1 \rho_{(x,y)}(x_i - \overline{x})\big(T'(y) - \overline{y}\big)dxdy = 0 \end{aligned} \tag{15}$$

The autocorrelation function ACF of the sequence is equal to the delta function.

Since the encryption object of privacy removal is A numeric quantity, the sequence $\{x_k\}$ composed of real numbers is a pseudo-random sequence formed by integers to achieve the purpose of privacy removal.

In order to make the sequence more random and improve its encryption rate, it is most appropriate to choose an interval M of 6 for random values. The value of Y_k is 3, 4, and 5 decimal places after the decimal point of X_k, which will enhance the resistance to selective plaintext attacks.

In order to improve the fitness of plaintext attacks, data processing is carried out on X_n to make the relationship between X_n and X_{n+1} more complex, and then prevent attackers from using a simple derivation process to solve the μ value. This article uses interval retrieval to place a numerical value between X_n and X_{n+1}, and the correlation between X_n and X_{n+1} is transformed into:

$$X_{n+1} = \mu \times (\mu \times x_n \times (1 - x_n)) \times (1 - \mu \times x_n \times (1 - x_n)) \tag{16}$$

Equation (13) is a quadratic equation with one variable. Add N values between X_n and X_{n+1}, then the correlation between X_n and X_{n+1} is a system of degree N with one variable. N μ solutions can be approximated by Newton stepwise approach method. At the same time, if N is large enough, a small change in X_n will cause X_{n+1} to make a change that affects x'_n. Set the value of x'_n to three digits behind the decimal point.

In order to avoid exposing the redundancy of plaintext information, encryption systems need to fully and uniformly use the ciphertext space, and chaotic sequences also need to clarify the statistical distribution state of the encrypted ciphertext data, that is, whether the plaintext data is completely covered up, which is also a key criterion for weighing the effectiveness of encryption methods. If the key sequence is completely random, then the encrypted ciphertext data follows a uniform distribution. Treat the plaintext and key sequence as a single byte data stream, and arbitrarily select a byte of plaintext m and key k. Assuming that the values of m are not equal, that is, the probability of a certain bit appearing as 0 or 1 is different, and the occurrence of a data bit appearing as 0 or 1 is an independent event. If the probability of 1 appearing in the i bit of the plaintext is p, the key sequence conforms to white noise under ideal conditions, that is, the probability of generating 0 or 1 is the same, both are 0.5. So the probability of 1 appearing in the i bit of ciphertext c after encryption is:

$$\begin{aligned} P_{(C_i=1)} &= P_{(m_i=1,k_i=0)} + P_{(m_i=0,k_i=1)} \\ &= P \times 0.5 + (1 - P) \times 0.5 = 0.5 \end{aligned} \tag{17}$$

Therefore, it can be seen that the probability distribution of encrypted ciphertext is symmetrical.

3.3 Privacy Data Security Detection Methods

Social network privacy data security risk assessment is to take an all-round evaluation and measurement of data risk. Privacy risk refers to the possibility of data leakage and subsequent negative impact.

In the set pair analysis theorem, the correlation number describes the same, different and inverse relation analytic expression of two sets which are correlated with each other, restricted and highlighted. Assuming that Z is a non-empty set sum, then the relation number of A=$\{< z, a_A(z), b_A(z), c_A(z) > |z \in Z\}$ obtained is:

$$\mu_A(z) = a_A(z) + b_A(z)i + c_A(z)j \tag{18}$$

In the formula, $a_A(z)$, $b_A(z)$, and $c_A(z)$ are the support, uncertainty, and opposition of element A belonging to F within Z.

The above equation contains the same, different, and opposite factors, so it is also called a ternary connection number. By conducting in-depth calculations on the uncertainty value b_i within the ternary connection number, the general form obtained is:

$$\mu = a + b_1 i_1 + b_2 i_2 + \cdots + b_n i_n + cj \tag{19}$$

Partial correlation number is a function that represents the change level of deviation trend in machine learning, reflecting the development trend and change of the homologous and different anticorrelation form. In the process of social network privacy data security risk assessment, homogeneity means that privacy risk assessment and expected standard risk are close to the same change pattern, that is, in the low risk area of privacy assessment; Equilibrium means that the gap between privacy risk assessment and expected standard risk level is slightly higher, and it is in the intermediate risk area of privacy assessment. The backlash represents the opposite state of privacy risk assessment and expected standard risk. In the high risk area of privacy assessment.

The scoring function of the five-element correlation number μ is expressed as:

$$S(\mu) = (a + b) - (d + e) \tag{20}$$

The exact function of the five element connection number μ is:

$$H(\mu) = a + b + d + e \tag{21}$$

The score function $S(\mu)$ and the exact function $H(\mu)$ are similar to the mean and variance in data statistics. The higher the score function is, the greater the data implied risk is. The lower the scoring function, the smaller the risk index. When the scoring function is equal, the higher the precision function is, the greater the data risk is, and vice versa. Therefore, this paper uses score function and precision function to carry out hierarchical prediction of hidden risks of social network privacy data.

Weight refers to the numerical value of the impact of each indicator to be evaluated during data privacy risk assessment. This article combines subjective and objective

weighting to design a minimum binary weighting method for the potential of five element partial connection numbers.

Regarding privacy risk attribute G_i, describe its r-order partial connection potential matrix as:

$$P_i^{(r)} = \begin{pmatrix} p_{11}^{(r)} & p_{11}^{(r)} & \cdots & p_{1n}^{(r)} \\ p_{21}^{(r)} & p_{22}^{(r)} & \cdots & p_{2n}^{(r)} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1}^{(r)} & p_{m2}^{(r)} & \cdots & p_{mn}^{(r)} \end{pmatrix} = \left(p_{ij}\right)_{nm} \tag{22}$$

For two unequal r-order partial relation number potential matrices $P_j^{(r)}$ and $P_k^{(r)}$, the deviation is described by $D_i^{(r)}(\omega)$, which is specifically denoted as:

$$D_i^{(r)}(\omega) = \left(\sum_{i=1}^{n} \sqrt{\left(P_{ij}^{(r)} - P_{ik}^{(r)}\right)^2}\right)^{\frac{1}{2}} \tag{23}$$

Equation (6) only represents the state where the key levels of weights are equal. In general, the weight key levels of data privacy risk indicators are not equal to each other. To understand the applicability of the evaluation results within the evaluation indicators composed of different weight vectors W_i, it is necessary to establish a minimum binary function:

$$\begin{aligned} &\min\ D^{(r)}(\omega) = \sum_{i=1}^{m} D_i^{(r)}(\omega) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \omega_j^{(r)} \sqrt{\left(P_{ij}^{(r)} - P_{ik}^{(r)}\right)^2}\right)^{\frac{1}{2}} \\ &s.t. \sum_{j=1}^{n} \omega_j^{(r)} = 1,\ \omega_j^{(r)} \geq 0,\ j = 1, 2, \cdots, n \end{aligned} \tag{24}$$

The Lagrange function is constructed to calculate the above formula, and the analytical formula of minimum h-partial weighting is obtained as follows:

$$\begin{matrix}(r) \\ j\end{matrix} = \frac{\left[\sum\limits_{i=1}^{m} \left(\sqrt{\left(P_{ij}^{(r)} - P_{ik}^{(r)}\right)^2}\right)^h\right]^{\frac{1}{h}}}{\sum\limits_{j=1}^{n} \left[\sum\limits_{i=1}^{m} \left(\sqrt{\left(P_{ij}^{(r)} - P_{ik}^{(r)}\right)^2}\right)^h\right]^{\frac{1}{h}}} \tag{25}$$

A machine learning based method for detecting privacy data security in social networks has been implemented.

4 Experiment

4.1 Experimental Design

According to the calculation steps of the method in this paper, Logistic is taken as an example to explore the real utility of chaotic sequence encryption, and carry out the simulation of power marketing data privacy encryption and decryption.

Figure 2 is a schematic diagram of the ASC code value distribution of ciphertext characters after two groups of slightly different original keys are successively encrypted for the same plaintext. The dotted line represents the plaintext, and the square dotted line indicates the ciphertext character, $\mu = 3.74$, which is obtained after encryption and the value of x_0 is 0.60001. As can be seen from the figure, the plaintext of the same text will change obviously when the key changes slightly, reflecting the sensitivity of the ciphertext to the key.

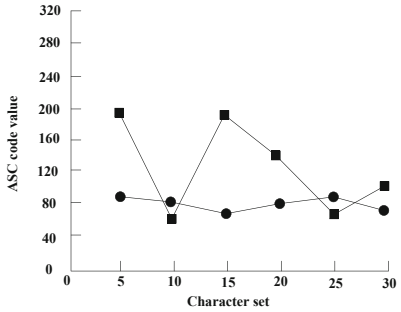

Fig. 2. ASC code value distribution diagram of ciphertext characters

Figure 2 shows the distribution of the number of iterations of the encryption method in this article. From the figure, it can be seen that the number of iterations is concentrated between 220 and 620, which reduces the number of iterations and improves the calculation speed and cost of the encryption method (Fig. 3).

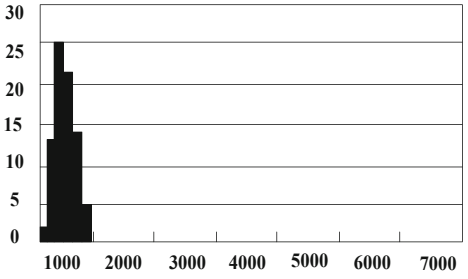

Fig. 3. Schematic diagram of the distribution of iteration times in this method

4.2 Experimental Results

In order to further verify the reliability of the security detection performance of social network private data, the proposed method is simulated and compared with literatures [3] and [4], and the security detection operation is conducted on the social network private data of an enterprise. The test environment is a Pentium 3.0GHzCPU with 2GB of memory. The three method performance pairs are obtained as shown in Table 3.

Table 2. Performance comparison of security detection methods for social network privacy data

Number of iterations	Social network privacy data security detection time/s		
	Method of reference [3]	Method of reference [4]	Textual method
1000	28.6	39.2	0.5
2000	35.1	48.9	0.9
3000	69.0	69.0	1.6
4000	92.3	78.1	2.9
5000	120.5	89.9	5.3
6000	138.9	125.3	8.1

From Table 2, it can be seen that when the number of iterations is 1000, the social network privacy data security detection time of the method in reference [3] is 28.6 s, the social network privacy data security detection time of the method in reference [4] is 39.2 s, and the social network privacy data security detection time of the method in this article is only 0.5 s; When the number of iterations is 5000, the social network privacy data security detection time of the method in reference [3] is 120.5 s, the social network privacy data security detection time of the method in reference [4] is 89.9 s, and the social network privacy data security detection time of the method in this article is only 5.3 s; The detection efficiency of the method proposed in this article is better than that of two references, which proves the superiority of the method in detecting privacy data security in social networks. This is because this method realizes the standardization of social Internet privacy data through N-Gram language model; The social Internet privacy data security risk detection based on the skew coefficient method in machine learning effectively reduces the time consumption of social Internet privacy data security detection, and improves the detection efficiency.

In order to further verify the security detection capabilities of different methods for social network privacy data, the accuracy of social network privacy data was tested, and the results are shown in Table 3.

According to Table 3, when the number of iterations is 1000, the accuracy of social network privacy data for the method in reference [3] is 66.2%, the accuracy of social network privacy data for the method in reference [4] is 72.1%, and the accuracy of social network privacy data for the method in this paper is 99.3%; When the number of iterations is 3000, the accuracy of social network privacy data for the method in reference [3] is 70.1%, the accuracy of social network privacy data for the method in reference [4]

Table 3. Precision of social network privacy data

Number of iterations	Social network privacy data accuracy/%		
	Method of reference [3]	Method of reference [4]	proposed method
1000	66.2	72.1	99.3
2000	68.9	76.5	98.6
3000	70.1	78.9	99.5
4000	73.8	68.5	96.0
5000	79.2	70.3	98.2
6000	66.0	69.2	99.1

is 78.9%, and the accuracy of social network privacy data for the method in this paper is 99.5%; The accuracy of the method presented in this paper for submitting network privacy data is much higher than other methods, indicating that the method presented in this paper has high data accuracy. This is because the method used in this article uses a semantic vector representation model to obtain topic semantic vectors, and matches the obtained topic semantic vectors to achieve accurate detection of privacy data.

5 Conclusion

This paper proposes a machine learning-based privacy data security detection method for social networks. Collecting privacy data of social network, constructing N-Gram language model to realize standardized processing of privacy data of social network; A representation model based on semantic vector is used to obtain the topic semantic vector, and the obtained topic semantic vector is matched. Finally, the partial contact number method based on machine learning is used to detect the security risk of social network privacy data. Experimental results show that the security detection time of social network privacy data under the proposed method is only 5.3 s, and the detection accuracy of social network privacy data is 99.5%, which proves the superiority of the proposed method for the security detection efficiency of social network privacy data.

References

1. Qiao, Y.: Research on social network security and privacy data fusion method based on cloud computing. Software **43**(06), 109–111 (2022)
2. Liu, C., Du, J., Zhou, N.: A cross media search method for social networks based on adversarial learning and semantic similarity. Chin. Sci. Ser. F **51**(05), 779–794 (2021)
3. Niu, N., Zhou, S., Lu, R., Yan, S., Zhang, M., Wang, C.: Attribute based signature encryption scheme based on cloud computing in medical social networks. J. Electron. Inf. Technol. **45**(03), 884–893 (2023)
4. Zhang, H.: Analysis of social network user data security protection based on big data background. Comput. Knowl. Technol. **18**(28), 69–71 (2022)

5. Zhu, P., Hu, J., Lv, S., et al.: Research on blockchain based privacy data protection methods for social networks. Inf. Sci. **39**(3), 94–100 (2021)
6. Jia, R., Wang, X., Fan, X.: Research on the influencing factors of personal information security and privacy protection behavior of social network users. Modern Intell. **41**(09): 105–114+143 (2021)
7. Wu, X., Liu, Q., Zhu, C.: Research on the application of collaborative public opinion fraud detection methods in social networks. J.Zhengzhou Univ. (Eng. Ed.) **43**(02), 7–14 (2022)
8. Zhu, D., Zhang, X., Gu, C.: Probability prediction of social network information leakage nodes based on EDLATrust algorithm. J. Tsinghua Univ. (Nat. Sci. Ed.) **62**(02), 355–366 (2022)
9. Zheng, Z., Wu, X., Wang, H., Liu, K., Shen, Z.: PTPM protection method for trajectory privacy in mobile social networks. Small Micro Comput. Syst. **42**(10), 2153–2160 (2021)
10. Lv, J., Zhang, Z., Xu, Y.: Blockchain based social network digital rights management protection method. Comput. Eng. Des. **42**(6), 1562–1570 (2021)
11. Li, Q., Hu, Y., Zhou, Q., Zhou, G.: K-anonymity based data privacy social network protection scheme. Modern Inf. Technol. **6**(09), 89–91 (2022)
12. Jianghui, F.: Data fusion method of social network security privacy based on cloud computing. J. Univ. Jinan: Nat. Sci. Ed. **35**(1), 5–16 (2021)
13. Zheng, J., Yang, L.: Large social network differential privacy algorithm based on Singular value decomposition. Comput. Technol. Dev. **32**(3), 126–131 (2022)
14. Zhang, Y., Zhang, J.: Research on the protection of social Internet privacy in MapReduce model based on K-means method. Wirel. Internet Technol. **19**(20), 162–165 (2022)
15. Zhu, Y.: A method to eliminate redundancies in browsing behavior data of social network users based on Random forest. J. Ningxia Normal Univ. **42**(1), 73–78 (2021)