A False Bank Accounts Detection Method Based on Naive Bayesian Network

Weiqing Wan*, Li Weihong

1 Jiangxi University of Engineering, Xinyu, Jiangxi, China

*Corresponding author e-mail: 200900031@hzxy.edu.cn

Abstract. At present, more and more cases of economic fraud are based on false bank accounts. Due to the limitation of the bank's technology and ability to verify the true identity of customers, the online verification system of banks can only check the authenticity of the ID card, while the consistent verification of people and cards can only be completed manually by bank tellers. It is easy to produce errors if the consistency is judged by the teller's manual naked eye. In this paper was proposed a method of constructing a naive Bayesian network classification model by automatically detect the authenticity of bank accounts to avoid the problem of economic fraud of false bank accounts, and verified the feasibility of this method through specific examples. The experimental results show that the method which has certain theoretical and practical significance, and provides reference for the detection of other false accounts.

1. Introduction

With the development of telecommunication network, the means of fraud continue to be renovated. Among all the economic frauds, it is basically through a very important crime tool - false bank account. After "telling the story of fraud", criminals use various realistic reasons to lure the victim to withdraw money from the false account, or swindle the other party's bank account password and dynamic verification code to transfer the funds in the card. The door to financial fraud was opened by fake accounts and the detection of such frauds has also become more difficult.

Wells Fargo has released an investigation and found that as many as 1 million additional savings users and bank card accounts may have been opened without the authorization or knowledge of employees, bringing the total number of false bank accounts to about 350. Wells Fargo also claimed that there were two-thirds more false bank accounts than had been estimated, posing serious economic risks.

There are not only some false bank accounts, but also many kinds of login accounts on the network. From application software, official website or network platform to real or false or automatic interaction with users, it is more and more difficult to effectively identify false accounts. Research shows that these false accounts can automatically generate complex comment information through artificial intelligence, and this false comment information cannot be detected by the machine, even the readers are not easy to distinguish.

Generally speaking, the means for banks to verify whether the customer's real identity is consistent with the account number is relatively limited. When fraudsters handle banking business with their own real ID cards, the bank networking system can only check the authenticity of ID cards, and the consistency of people and cards can only be done manually by bank tellers, but it is easy to misjudge only by manual checking by tellers. With the promotion and use of bank self-service card opening machines, it is more difficult to verify the identity of customers.

This paper puts forward the method of constructing naive Bayesian network classification model to
automatically detect false bank account number, and automatically identify false bank account number to solve the problems of various kinds of false bank fraud, which also has certain reference significance for the false identification of other accounts.

2. Materials and Methods

2.1. Naive Bayes Classifier

Naive Bayes classifier is a series of simple probability classifiers based on Bayes theorem, which is based on the strong (naive) independence between the assumed features\[^1\]. The classifier model assigns the class labels represented by Eigen values to the problem instances, and the class labels are taken from a finite set. Due to the assumption of independent variables, it only needs to estimate each variable. One advantage of naive Bayes classifier is that it only needs to estimate the necessary parameters (mean and variance of variables) based on a small amount of training data.

Among all the automatic classification mathematical models, the naive Bayesian classification model is the simplest and efficient\[^2\]. It only assumes that for the characteristic variable \(x\), each feature \(X_i\) in the sample data is independent and identically distributed. The classification result is judged and detected by probability reasoning. The theoretical basis is Bayesian theorem in probability theory and naive Bayesian network, which has the advantage of stable classification performance\[^3\].

Naive Bayes classifiers only make strong assumptions about conditional probability \(p(X = x|Y = c)\), as shown in formula (1), independent and identically distributed\[^4\]. The conditional probabilities of each characteristic variable \(x\) and \(y\) are irrelevant, that is, there is \(P(AB) = P(A) P(B)\).

\[
P(X = x | Y = c_k) = P(X^{(i)} = x_1, \ldots, X^{(n)} = x_n | Y = c_k) = \prod_{j=1}^{n} P(X^{(j)} = x_j | Y = c_k)
\]

From the posterior probability, the value of \(P(Y = c_k | X = x)\) can be calculated, as shown in formula (2) and formula (3).

\[
P(Y = c_k | X = x) = \frac{P(X = x | Y = c_k) P(Y = c_k)}{\sum_k P(X = x | Y = c_k) P(Y = c_k)} = \frac{P(Y = c_k) \prod_j P(X^{(j)} = x_j | Y = c_k)}{P(Y = c_k) \prod_j P(X^{(j)} = x_j | Y = c_k)}
\]

\[
k = 1, 2, \ldots, K
\]

There is an ideological basis of naive Bayes: For a given item to be classified, which category is considered to belong to by solving the probability of each category appearing under the condition of this item appearing. Naive Bayes classifier theory is simple, assuming that each feature variable \(X_i\) is independent and identically distributed. Compared with Bayesian network, the parameters in the model are easy to obtain, and the classification results are similar to the characteristics of Bayesian network, so it has inherent advantages in the detection of false bank accounts\[^5\].
2.2. Naive Bayes Classification Method

According to naive Bayes theory, suppose that a sample \( X \) has \( n \) features \( X_1, X_2, \ldots, X_n \), which can be divided into \( m \) categories \( C \), expressed as \( C_1, C_2, \ldots, C_m \). Bayesian classifier generally calculates the probability of sample \( x \), then the category \( C \) with the largest probability is classified, that is to say, to find the maximum value of the following formula (4):

\[
P(C \mid X) = \frac{P(x_1, x_2, \ldots, x_n \mid C)P(C)}{P(x_1, x_2, \ldots, x_n)}
\]  

(4)

Since the joint probability \( p(x_1, x_2, \ldots, x_n) \) of \( N \) features \( X_i \) is consistent for all classes \( C \), it can be omitted in the calculation process, and only the maximum value of its conditional probability \( P(x_1, x_2, \ldots, x_n \mid C) \) \( P(C) \) can be calculated. It is obviously different from the ordinary Bayesian classifier. In practical application, in order to calculate conveniently, the above formula can be simplified as follows:

\[
P(C \mid X) \propto P(x_1, x_2, \ldots, x_n \mid C)P(C) = P(x_1 \mid C)P(x_2 \mid C)\ldots P(x_n \mid C)P(C)
\]  

(5)

Here, \( P(x_i \mid C) \) is the prior probability of characteristic \( X_i \), which can be obtained from historical log data. From the above formula, we can calculate the probability that sample \( x \) should classify each category \( C_1, C_2, \ldots, C_m \), so as to find the class \( C_k \) with the maximum probability. Although the hypothesis that "all features \( X_i \) are independent and identically distributed" is unlikely to be true in real life, it can greatly simplify the calculation of probability, and some studies show that it has little influence on the accuracy of classification results, which is very conducive to classification detection.

The prior probability \( P(x_i \mid C) \) and class probability \( P(c) \) in naive Bayes classifier can be obtained by learning and calculating the parameters of large samples, and only need to train and estimate the values of \((1 \leq i \leq n, 1 \leq k \leq m)\), so that the characteristic variable \( X = x_i \) can be classified as \( y = C_k \):

\[
P(Y = c_k \mid X = x) = \frac{P(X = x \mid Y = c_k)P(Y = c_k)}{\sum_k P(X = x \mid Y = c_k)P(Y = c_k)}
\]  

(6)

After observation and calculation of large sample categories, as shown in formula (7) and formula (8).

\[
P(Y = c_k) = \frac{s_k}{s}
\]  

(7)

\[
P(X^{(j)} = x_j \mid Y = c_k) = \frac{s_{kj}}{s_k}
\]  

(8)

Set \( s_k \) is the number of training samples in the training samples. Therefore, the values of category probability and prior probability can be obtained through parameter learning of large samples.

Vividly, the naive Bayesian network classifier can be as shown in Figure 1, which is conducive to detecting the falsity of bank account number by obtaining \( n \) features \( x_n \) of bank account number.

\[
\begin{align*}
n \text{ characteristic} \\
\uparrow \\
x_1 & \quad x_2 \quad \ldots \quad x_{n-1} \quad x_n \\
P(C \mid X) & \quad P(C) \\
\end{align*}
\]

Figure 1 Naive Bayesian Network Classifier
3. Results & Discussion

In this chapter, an example is given to illustrate the process of constructing naive Bayes classifier to detect false bank accounts. According to the sampling statistics of a community website, only 89% of the historical accounts of the website are real accounts (set as category \( C_0 \)), and 11% of the historical accounts are false accounts (set as category \( C_1 \)).

Suppose that a bank account number (sample \( X_i \)) has the following three characteristics \( (X_1, X_2, X_3) \), as shown in Table 1.

| Characteristic  | Value  |
|----------------|--------|
| \( X_1 \): number of logs / days of registration | 0.1 |
| \( X_2 \): number of friends / registration days | 0.2 |
| \( X_3 \): do you want to use a real picture | 0 |

From the above historical data, based on Naive Bayes classifier, the authenticity of an account can be judged by the prior probability \( p(X|C) \) of historical statistical data, and the value can be calculated, as shown in formula (9).

\[
P(C \mid X) \propto P(x_1 \mid C)P(x_2 \mid C)\ldots P(x_n \mid C)P(C) \tag{9}
\]

The above Eigen values \( X_i \) can obtain their prior probabilities from historical statistics, but the random variables \( X_1 \) and \( X_2 \) are continuous, which are not suitable for the classification calculation of naive Bayesian model. Therefore, it is necessary to discredit the continuous random variable \( x \) and replace the probability of the continuous random variable with the discrete probability. According to Gaussian probability distribution, if \( x < 1 \) is decomposed into three continuous intervals \([0, 0.05]\), \((0.05, 0.2)\) and \([0.2, +\infty)\), the probability of each interval is calculated respectively.

The Gaussian distribution function can be expressed as shown in formula (10).

\[
f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} \tag{10}
\]

Then the probability of \( X \) interval \((a, b)\) of random variable can be calculated as shown in formula (11).

\[
P(x) = \int_a^b f(x)dx \tag{11}
\]

For example, since the random variable \( X_1 \) is equal to 0.1, replace it with the probability \( P(x_1) \) falling in the second interval \((0.05, 0.2)\), as shown in Figure 2.

![Figure 2 Discretization of Continuous Random Variables](image-url)
Therefore, the probability of classification into $C_0$ (normal use account) and $C_1$ (abnormal use account) is calculate, as shown in formula (13) and formula (14).

\[
P(C_0 | X) = P(x_1 | C_0) P(x_2 | C_0) P(x_3 | C_0) P(C_0) / P(x_1, x_2, \ldots, x_n)
\]

\[
= 0.5 \times 0.7 \times 0.2 \times 0.89
\]

\[
= 0.0623 / P(x_1, x_2, \ldots, x_n)
\]

\[
P(C_1 | X) = P(x_1 | C_1) P(x_2 | C_1) P(x_3 | C_1) P(C_1) / P(x_1, x_2, \ldots, x_n)
\]

\[
= 0.1 \times 0.2 \times 0.9 \times 0.11
\]

\[
= 0.00198 / P(x_1, x_2, \ldots, x_n)
\]

Here, $P(x_1, x_2, \ldots, x_n)$ is a positive real constant. Therefore, although the user does not use a real avatar, the probability that he is a real account is more than 30 times higher than that of a false account. Therefore, it can be judged that the account is true.

4. Conclusions
With the development of the Internet, more and more cases of economic fraud are based on false bank account numbers, which are provided to criminals. As the means of verifying the true identity of customers by banks is limited, the online verification system of banks can only query the authenticity of identity cards, while the consistent verification of people and cards can only be completed by tellers manually, which is easy to be generated by the manual judgment of tellers Misjudgment. This paper proposes to build a naive Bayesian network classification model to automatically detect the falsity of bank accounts to solve the problems that the current bank customer identity false economic fraud. The experimental results show that the method which has certain theoretical and practical significance, and provides reference for the detection of other false accounts.

Acknowledgements
This paper supported by the research subject on the Teaching Reform of Colleges and Universities , Jiangxi Province: Research and Practice of "1+2+1" Application-oriented Talent Training Mode under the Background of New Engineering Transformation -- A Case Study of Internet of Things Engineering (JXJG-19-28-3), and the Industrial Support Project of Xinyu Science and Technology Bureau, Jiangxi Province: Research on Reliability Control Technology for Train Communication Network Transmission System (2019).

Author
Weiqing Wan (1973.10-), male, Fengcheng county, Jiangxi Province, Han, postgraduate degree, associate professor, majoring in: software engineering, trusted software. Li Weihong (1984.03-), female, LouDi city, Hunan province, Lecture.

References
[1] Yao Peijin. Research on Integrated Algorithms Based on Naive Bayes[D]. Guangzhou: South China University of Technology, 2013.
[2] Ma Maoyuan, Lu jia. Naive Bayes Semi-Supervised Self-Training Method Combine with Similarity[J], Journal of Chongqing Normal University (Natural Science), 2019, 36(1):78-84.
[3] Sun Mei, Zhang Sen, Nie Peiyao, Nie Xiushan. Research on session segmentation of web search...
query logs based on naive Bayes[J], Journal of Nanjing University (Natural science), 2018, 54(6): 1132-1140.

[4] Zhang Wei, Wang Zhihai, Yuan Jidong, Liu Haiyang. A locally attribute weighted naive Bayes classifier[J], Journal of Beijing jiaotong University, 2018, 42(2): 14-21.

[5] YANG Lei, CAO Cui-ling, SUN Jian-guo, ZHANG Li-guo. Study on an improved naive Bayes algorithm in spam filtering[J], Journal on Communications, 2017, 38(4): 140-148.