Detecting Bank False Account Based on Naive Bayesian Network

Weiqing Wan, Qingyan Zeng and Zhicheng Wen*
Jiangxi University of Engineering, Xinyu, Jiangxi Province, 338000, China

*Corresponding author e-mail: zcwen@mail.shu.edu.cn

Abstract. With the development of Internet, more and more fraudulent cases are committed on the basis of bank false accounts, but the means and ability of bank to verify the true identity of customers are limited. Generally, a bank online verification system can only check the authenticity of identity cards through inquiry, while the verification of people and cards can only be completed manually by Bank tellers, but only by the counter. Personnel's naked eye judgment is prone to potential risks. Aiming at the problem of fraud of bank customer identity, this paper proposes a method of automatic detecting bank account fraud using Naive Bayesian Network classification model, and verifies the feasibility of this method through specific examples.

1. Introduction

At present, the hands of telegraph fraud are new and type frequently, and there are fake bank accounts that are very important crime tools in all fraud cases. After the premise is laid out after the premise is made, the criminal and the verification code of the bank account are cheated by the criminal, and the fund is sent every time the premise is laid out. If the account that received the fraud fund is a false account, the crime for the fraud of the criminal is opened, and the police is increasing the difficulty of solving the case.

U. S. Wells Fargo issued a statement Thursday. According to an external survey, even more than 1.40 million savings and credit card accounts could be opened in the absence of employees when the bank encourages retail customers to sell more goods. According to the U. S. wealth bank, the false accounts made by the company's employees are two thirds more than the initial estimate, and the false account scandals have brought a record fine to the bank and invited the investigation of the diet.

There is a certain falsehood in the account on the net as well as the falsehood of the bank account. From creating applications software, official sites, or content diffusion platforms to producing images, videos, or texts with substantial content, or until fake or automatically By interacting with the false accounts, the accounts seem increasingly "true". There is a research result that can generate complex comment information to artificial intelligence, but these false comments are not understood only to machine but human reader.

In general, there are limited ways for banks to identify customers' true positions. When the criminal processes the business with the real identification, the online collation and examination system of the bank can examine only the truth of the identification card, and the collation and examination of the person and the certificate can be completed only by the bank. However, it relies only on the naked eye of the window, and decides that it is easy to generate the danger of hiding the risk. Over the past few
years, it has become difficult for customers to confirm their identity on the bank's customers with the spread of self-card purchases. Therefore, the situation of the bank's false account is severe.

For the problem of false fraud of the present position of the bank in the bank, the body uses the model of the classification of naive Bayesian network to examine and measure the falsehood of the bank's account number, and identify the account of the false bank itself, and a certain theory And the significance of practice.

2. Naive Bayesian Classifier

Among all the classification models, Naive Bayesian classification model is simple and efficient. It is only assumed that for the characteristic variable X, the components Xi in the sample data are independent and identically distributed, and can be judged and detected by probability reasoning. Its theoretical basis is Bayesian theorem and naive Bayesian network in probability theory, which has the greatest advantage of stable classification performance.

In all classification models, Naive Bayesian classification models are simple and efficient. Assuming that only the characteristic variable x is assumed, each component Xi of the sample data is the same distribution independently of each other, and it is judged and judged by the probability theory, and the theory is based on the Bayesian theorem and naive Bayesian network theory in the theory of probability, and the greatest advantage is It has stable classification performance.

The Naive Bayesian classification model is generally not very sensitive for missing data samples, and the complexity of time is O(m*n), and the classification algorithm can be realized easily. Sample X input to Naive Bayesian classifier is assumed to be a set of n-dimensional vectors, while the output is a set of category labels Y={c1,c2,...,ck}. When an input n-dimensional vector x∈X is input, the classifier outputs the category label y∈Y to which it belongs. Where x and y are random variables on sets X and Y, respectively. The large sample training set of the classifier T={(x1,y1), (x2,y2),..., (xn,yn)}is generated by the independent and identical distribution of the joint probability distribution P (X, Y).

Naive Bayesian classifier makes a strong assumption on conditional probability P (X = x | Y = ck), that is, conditional independence assumption. The conditional probability of each characteristic variable is independent and identically distributed. We have:

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

From Bayesian theorem in probability theory, the value of P (Y= c_k|X = x) can be calculated.

\[ P(Y = c_k | X = x) = \frac{P(X = x | Y = c_k)P(Y = c_k)}{\sum_k P(X = x | Y = c_k)P(Y = c_k)} \]
\[ = \frac{P(Y = c_k)\prod_j P(X^{(j)} = x_j | Y = c_k)}{\sum_k P(Y = c_k)\prod_j P(X^{(j)} = x_j | Y = c_k)}, k = 1,2,\ldots,K \]

Naive Bayesian classifier theory is simple and clear. Assuming that each characteristic variable xi is independent and identically distributed, the parameters in the model are easy to obtain and the classification results are similar to those of Bayesian network. It has inherent advantages in false detection of bank accounts.

The Naive Bayesian classifier theory is simple and clear. It is easy to obtain the parameters in the model if the independent distribution between each characteristic variable xi is assumed, and the result of the classification is similar to the feature of Bayesian network. It has inherent advantages in false detection of bank accounts.
3. Naive Bayesian Classifications

According to the Naive Bayesian theory in Chapter 2, assuming that a sample X has n features, it can be expressed as x1, x2,... And xn, and can be differentiated into m categories C (Category), C1, C2,... and Cm. The general Bayesian classifier generally calculates the classification with the highest probability of category C of sample X, that is, the maximum of the following formula:

\[ P(C|X) = \frac{P(x_1x_2...x_n|C)P(C)}{P(x_1x_2...x_n)} \]

Since the joint probability \( P(x_1x_2...x_n) \) of n characteristic \( Xi \) is the same for all category \( C \), it can be omitted in the calculation process, so only the maximum of conditional probability \( P(x_1x_2...x_n|C)*P(C) \) is required. The naive Bayesian classifier assumes that all features \( Xi \) are distributed independently and identically in order to calculate conveniently. Therefore, the above formula can be simplified as follows:

\[ P(C|X) \propto P(x_1x_2...x_n|C)P(C) = P(x_1|C)P(x_2|C) ... P(x_n|C)P(C) \]

\( P(x_i|C) \) is the prior probability of characteristic \( x_i \), which can be obtained from historical statistics. From this, we can calculate the probability that sample X should belong to each category C, so as to find the class with the maximum probability. Although the hypothesis of "all features Xi are distributed independently and identically" is unlikely to hold true in reality, it can greatly simplify the calculation of probability, and some studies show that it has little impact on the accuracy of classification results, which is very conducive to the detection of bank account fraud.

The priori probability and class probability of Naive Bayesian classifier can be obtained by machine learning the parameters of large samples. Only the training value \( P(Y=c_k) \) and \( P(X^{(i)}=x_j|Y=c_k) \) (1 < i < n, 1 < k < m) are needed. Thus the characteristic variable \( X = x_i \) can be classified as \( Y = c_k \):

\[ P(Y = c_k | X = x) = \frac{P(X = x | Y = c_k)P(Y = c_k)}{\sum_{k} P(X = x | Y = c_k)P(Y = c_k)} \]

Here, after a large sample observation, there are

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

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

Here, \( s_k \) is the number of samples with \( c_k \) in the training set, \( s \) is the total number of training sets, and \( s_{kj} \) is the number of samples with \( c_k \) in the training set and \( x_j \) in the attribute value. Therefore, the value of sum \( P(Y=c_k) \) and \( P(X^{(i)}=x_j|Y=c_k) \) can be obtained through parameter learning of large sample.

Imaginably, the Naive Bayesian Network classifier can be shown in Figure 1, which is helpful to detect the falsity of bank accounts by acquiring n features of bank accounts.

4. Examples of false account detection

In this chapter, an example is given to illustrate the process of using Naive Bayesian classifier to detect bank false accounts. According to the sample statistics of a community website, only 89% of the 10,000 accounts of the website are real accounts (set as category C0) and the other 11% are false accounts (set as category C1).
Assuming that a bank account (sample) has three characteristics (x1, x2, x3), as shown in the following table:

| Characteristic X | x1: Number of logs /days of registration | x2: Number of Friends /Days of Registration | x3: Whether to use real avatars or not |
|------------------|------------------------------------------|---------------------------------------------|---------------------------------------|
| Characteristic   | 0.1                                      | 0.2                                         | 0                                     |

Starting from the above data, using naive Bayesian classifier and the prior probability of historical statistics, we can judge the authenticity of an account. We can mainly calculate the value of the following formula.

$$P(C|X) \propto P(x1|C)P(x2|C)P(x3|C)P(C)$$

These characteristic values can get their prior probability from statistical data, but the characteristic x1 and x2 are continuous random variables, which are not suitable for the probability calculation of Naive Bayesian classification model. It is necessary to discretize the continuous random variables and replace the probability of discrete random variables with the probability of continuous intervals. If x1 is decomposed into three continuous intervals $[0, 0.05]$, $(0.05, 0.2)$ and $[0.2, +\infty]$, the probability of each interval is calculated. Since x1 equals 0.1, the probability of falling into the second interval is replaced.

According to historical statistics, the prior probability can be obtained:

- $P(x1|C0) = 0.5$, $P(x1|C1) = 0.1$
- $P(x2|C0) = 0.7$, $P(x2|C1) = 0.2$
- $P(x3|C0) = 0.2$, $P(x3|C1) = 0.9$

Hence,

- $P(C|X)=P(x1|C0) P(x2|C0) P(x3|C0) P(C0) / P(x1x2...xn)$
- $= 0.5 * 0.7 * 0.2 * 0.89 = 0.0623/ P(x1x2...xn)$
- $P(C|X)=P(x1|C1) P(x2|C1) P(x3|C1) P(C1) / P(x1x2...xn)$
- $= 0.1 * 0.2 * 0.9 * 0.11 = 0.00198/ P(x1x2...xn)$

Here, $P (x1x2... xn)$ is a positive real number. As you can see, although this user does not use a real avatar, he is more than 30 times more likely to have a real account than a false one, so you can judge that the account is true.

5. Conclusions

According to the development of the Internet, fraud has increased based on bank accounts. It is convenient for the criminal, but the means that the bank confirms the true position of the customer is limited. Judging from the naked eye of the attendant, dangerous hidden danger is likely to occur. Therefore, in this paper, we have proposed to automatically detect the falsehood of bank accounts using a simple beetle net classification model for the problem of the imposition of customers in the present bank.

6. Acknowledgements

Weiqing Wan(1973.10-), male, Fengcheng county, Jiangxi Province, Han, postgraduate degree, associate professor, majoring in: software engineering, trusted software. Qingyan Zeng(1974.10-), male, Wan'an county, Jiangxi Province, Han, postgraduate degree, lecturer, majoring in: network Engineering. Corresponding author: Zhicheng Wen (1972.11-), male, Dong’an county, Hunan province, Han, Ph.D., professor, majoring in: network security, trusted software.
References

[1] P. Yao, Research on Integrated Algorithms Based on Naive Bayes, Guangzhou: South China University of Technology, 2013.

[2] M. Ma, Lu jia, Naive Bayes Semi-Supervised Self-Training Method Combine with Similarity, Journal of Chongqing Normal University(Natural Science), 36(1)(2019) 78-84.

[3] M. Sun, S. Zhang, P. Nie, X. Nie, Research on session segmentation of web search query logs based on naive Bayes, Journal of Nanjing University(Natural science), 54(6)(2018) 1132-1140.

[4] W. Zhang, Z. Wang, J. Yuan, H. Liu, A locally attribute weighted naive Bayes classifier, Journal of Beijing jiaotong University, 42(2)(2018) 14-21.

[5] L. Yang, C. Cao, J. Sun, L. Zhang, Study on an improved naive Bayes algorithm in spam filtering, Journal on Communications, 38(4)(2017) 140-148.