Research and Improvement of Internet Financial Anti-Fraud Rules Based on Information Gain and Support

Xiaoguo Wang, Zhichao Yi and Hao Wu

College of Electronics and Information Engineering, Tongji University, Shanghai, China
Email: yizhichao@tongji.edu.cn

Abstract. Using rules to develop internet financial anti-fraud is one of the important means in this field. How to improve the existing rules and give better performance to their effectiveness is a hotspot in the field of internet financial anti-fraud. Based on the FOIL information gain, Gini coefficient, and support degree, this paper investigates how to improve the internet financial anti-fraud rules with the combination of rules learning and improving technology. With the combination of these indexes, this paper analyzes the experimental results with using the real transaction data of banks. The experimental results show that the improved rules have good performance in support, and have some improvement for the internet financial anti-fraud.

1. Introduction
In the field of internet financial anti-fraud, Researchers are trying to use Convolutional neural networks [1], Feature engineering strategies [2], Cost-sensitive decision tree approach [3], Bayes minimum risk with cost sensitive [4] and Hidden Markov Model [5] to detect fraud transactions. Nowadays in the field of the application in the internet financial anti-fraud, People are some deficient in the respect of improving anti-fraud rules. This paper combines FOIL information gain [6] and rule support [7] as a measure of the rule effectiveness, makes research and improvement on rules of internet financial anti-fraud.

2. Related Work
In the field of rule engine and classifier based on rules, they use IF-THEN rule to classify. The "IF" part of rule is called the prerequisite. The rule prerequisite is composed of one or more logical connectives. The "THEN" part of a rule is called the result or conclusion of rules. The conclusion of a rule includes prediction of a class or possibility prediction of classes. It is similar to the Decision Tree [8], but rules can clearly express the information contained in it. For a given transaction, we call it a tuple. If the conditions in the rule precursor are all set up, then we say that the rule precursor is satisfied and the rule covers the tuple, that is, the rule covers the transaction.

2.1. Rule Evaluation
Rule has two classic important indicators -- coverage and accuracy. Now given a set of internet financial transaction data records, it includes P pen fraud transactions and N pen non-fraud transactions. For a specific rule $R_1$, $p_1 + n_1$ is the total number of transactions covered by the rule, in which $p_1$ is the number of transactions that are correctly classified, $n_1$ is the number of transactions for the wrong classification, then we have:
The rule coverage is the percentage of the transactions covered by the rules, and the accuracy of the rule is the percentage of the correctly classified transactions covered by the rules. On the other hand, in the field of association rule [7, 9], rules are evaluated by their support and confidence. The expression of support and confidence is as follows:

\[
\text{support}(R_i) = \frac{p_i}{P + N} \quad (3)
\]

\[
\text{confidence}(R_i) = \frac{p_i}{p_i + n_i} \quad (4)
\]

It can be seen that confidence is the same with the accuracy, and support is the proportion of the right classified transactions in all transactions. It can be see that the support is more inclined to the higher accuracy than the coverage.

2.2. FOIL Information Gain and Gini Coefficient

Foil information gain is an evaluation proposed from FOIL algorithm [6]. FOIL is a covering algorithm for learning first order logic rules. In the classification algorithm based on rules, the transactions satisfying the rules called positive tuples, the rest of the tuple called negative tuples. The FOIL algorithm determines the validity of the rule by comparing the information gain of original rules and conjunctive rules, and determines whether the original rules should be replaced by the conjunctive rules. The expression of attribute \( v \) information gain (\( \text{Gain}(v) \)) and rule \( R_i \) information gain (\( \text{Gain}(R_i) \)) is as follows:

\[
\text{Gain}(v) = v_1(\log_2 \frac{v_1}{v_1 + n_1} - \log_2 \frac{P}{P + N}) \quad (5)
\]

\[
\text{Gain}(R_i) = p'_1(\log_2 \frac{p'_1}{p'_1 + n'_1} - \log_2 \frac{p_1}{p_1 + n_1}) \quad (6)
\]

Among it, \( v_1 \) is the number of fraudulent transactions covered by the attribute \( v \) and \( n_1 \) is the number of non-fraudulent transactions covered by it. \( P \) is the number of fraudulent transactions in all transactions, and \( N \) is the number of non-fraudulent transactions. Meanwhile for rule \( R_i \), \( p'_1 \) is the number of fraudulent transactions covered by the improved rules, and \( n'_1 \) is the number of non-fraudulent transactions. \( p_1 \) is the number of fraudulent transactions covered by the original rules, and \( n_1 \) is the number of non-fraudulent transactions.

In the current situation of internet financial transactions, the number of fraudulent transactions is far less than that of non-fraudulent transactions. So \( P \) is far less than \( N \), which makes the value \( \log_2 \frac{P}{P + N} \) in the information gain tend to be negative infinity. As a result, the information gain value will depend on the value of the number of fraudulent transactions covered by the attribute. And it cannot determine whether the characteristic attribute will be inclined to fraudulent transactions. On the other hand, as regard to rule information gain it always tends to improve the accuracy and confidence of rules to a very high value, which will lead to a small support value, and the quality of the rules is not high.

In order to solve these two problems, we combine the information gain and the Gini coefficient to select better characteristic attributes. Gini coefficient is a feature selection method similar to attribute
information entropy. Classification and Regression Tree (CART) [10] algorithm uses Gini coefficient to construct binary decision tree. For an attribute \( v \), its Gini coefficient is calculated as follows:

\[
Gini(v) = \sum_{i=1}^{n} p_i (1 - p_i) = 1 - \sum_{i=1}^{n} p_i^2
\] (7)

Among it \( p_i \) is the probability of the class \( i \) when attribute \( v \) is determined. In the internet financial anti-fraud, we make an assumption that when the attribute \( v \) is determined, all of the transactions are fraudulent transactions. So there are \( p_i = 1 \), therefore \( Gini(v) = 0 \). In this situation, the transactions are biased towards fraudulent transactions. It can be seen that the smaller the Gini coefficient of attribute \( v \) is, the lower uncertainty of data is. In the feature selection, we choose the attributes that its Gini coefficient are not higher than certain values. Then we select some of the attributes whose information gain is higher and make them conjunct into the rules.

In order to solve the problem that low quality of the improved rules, we combine the two criteria of FOIL: information gain and rule support to determine whether the rule is more effective than the original.

3. Research and Improvement Based on Information Gain and Support on Rules of Internet Financial Anti-Fraud

3.1. Data Preprocessing

There are two main characteristics of internet financial transaction data: Firstly, data size is large; Secondly, varied transaction types lead to diverse transaction data. These unprocessed data are difficult to adapt to the internet financial anti-fraud rule engine and machine learning model. Good data samples have significantly improved the learning of the model and the execution of the engine. For data pre-processing, we divide data into two parts: original variable and statistical variable.

Original variables: generally, it refers to the information of the current transaction, such as the amount of transaction, the type of transaction, the trading channel, the place of the transaction, the transaction user, the transaction time, etc. For the original variable, firstly we must handle the missing value and extreme value, and the variable which is missing too seriously may need to be rounding out or regard the missing value as a special attribute value of the variable; and then we the basic transaction information is explored, for example, compressing the continuous variable (transaction amount and so on).

Statistical variables: the attributes of the internet financial transactions are few, so directly applying them to the rule engine is not ideal. Therefore, statistical variables should be produced by the method of statistics, in which there are statistical variables based on time. For example, the total amount of transaction money of a user in 30 minutes. And there are the statistical variables based on events. For example, the total money of \( n \) pen transaction in the past. Using these extended data variables to provide more composition of the feature selection.

3.2. Research Method

Most traditional classifiers based on rules use top-down strategies. It gradually improves rule by add or replace attribute from a normal rule until it meets a predetermined rate of accuracy. It is a process of gradual specialization, from normal to special. Corresponding to it is the bottom-up process, gradually improving rules from special to normal. However, internet financial transactions have a large amount of data. It is time-consuming and hard to improve rules from the very beginning. And the coverage rate from special to normal rules is not ideal. So we can build an internet financial fraud rule data base by analysing historical data and experience speculating. And improve the rule base so that it can reach our predetermined target. At the same time, in order to solve the problem that the influence of attribute coverage in attribute information gain is far greater than the accuracy, we do not select the maximum gain value for each rule. But we consider the Gini coefficient of the attribute, and select some attribute values that their Gini coefficient is not greater than a certain value. Determining which is more
effective by the information gain and support of these improved rules. The following is the method procedure used in this paper.

First of all, assuming that we have a Dataset \( D = P \cup N \) and a rule data set \( R \). \( P \) is a fraudulent transaction in data set, \( N \) is a non-fraudulent transaction in data set. For each rule \( r_i \) in our rule set, we select the feature \( c_1,c_2,\ldots,c_m \) whose gini coefficient is not greater than \( \max \text{gini} \) meanwhile their information gain is as big as possible from condition library \( C \) and pop the \( c_1,c_2,\ldots,c_m \) from \( C \). Then we append or replace the feature \( c_1,c_2,\ldots,c_m \) for rule \( r_i \), so we have several unhandled improved rules: \( r_i', r_i'', \ldots, r_i'^m \). So our purpose has changed into finding the most effective rule in rule set: \( r_i'^1, r_i'^2, \ldots, r_i'^m \).

We calculate each rule’s information gain and its support value. Through this rule set we build a new rule set \( r_{i\_temp} \) that each rule in it stratifies following conditions: its support value is not less than the original rule’s support value and its information gain is not less than a value depending on original rule’s information gain and minimum loss coefficient \( \text{Foil\_Loss} \). Then we calculate the confidence and support for each rule in \( r_{i\_temp} \) and find the rule \( r_{i,k}' \) with biggest support value and its confidence is not less than minimum confidence in \( r_{i\_temp} \). Replacing the original rule \( r_{i} \) by improved rule \( r_{i,k}' \). We continue improving rule \( r_{i} \) until condition library \( C \) is empty or there is no condition matching rule \( r_{i} \) in condition library \( C \). Then we do above procedure for each rule in original rule set \( R \).

In the end, we compared the accuracy and support of the improved rule set and the original rule. If the accuracy rate is higher and the support loss is less than a certain value, the original rule set will be replaced by the improved rule set. Furthermore, we also do compared experiments that one is only using the information gain to select feature and better rule and another one is using the information gain, gini coefficient to do feature select and rule select.

4. Experiment and Result Analysis

4.1. Data and Experimental Method

The experimental data in this paper uses real transaction data of the cooperative bank, and table 1 illustrates the data used in this experiment, in which \( P \) represents suspicious fraud transactions and \( N \) represents non-fraudulent transactions.

| Table 1. Bank Real Transaction Data |
|------------------------------------|
| Total transaction                  |
| P       | N          |
| 263     | 111435     |

Due to the low confidence of the detection results in the internet financial anti-fraud, basing on the analysis of the bank’s real transaction data and combining with the experience of professionals in this field, the minimum requirement for the confidence degree of the rules is set as 0.06. At the same time, in the rule improvement process of the characteristic attributes chosen, the maximum Gini coefficient of the characteristic attribute \( v \) is set as 0.095. On this basis, the rule with the highest support is selected. Considering the large amount of transaction data for internet finance, it will take too much time to improve the rules. This experiment will use only 3 rules to make a test.

In this paper, the experiment is carried out distinguish by only using the FOIL information gain, combining the FOIL information gain and the Gini coefficient, combining the information gain, the Gini coefficient and the support degree as the index. Comparing the result to see the improvement of the rule.
Experimental environment: Linux system, 2GHz CPU, python 3.7.

4.2. Result Analysis
In this paper, the experiment is carried out on the real data test set with different indexes and methods, and the results of the final improvement are compared. The experimental results are shown in Table 2.

| Table 2. Experiment Result | Rule 1 | Rule 2 | Rule 3 |
|-----------------------------|--------|--------|--------|
|                             | P      | N      | P      | N      | P      | N      |
| Origin                      | 224    | 29049  | 163    | 6986   | 160    | 8196   |
| FOIL                        | 113    | 1914   | 53     | 420    | 52     | 661    |
| FOIL +Gini                  | 113    | 1914   | 53     | 420    | 52     | 661    |
| FOIL+Gini+ Support          | 113    | 1914   | 76     | 1026   | 90     | 1165   |

Among it, the Origin means the classification results by the rules without improving. P represents suspicious fraud transactions and N represents non-fraudulent transactions. FOIL means only using the information gain to select feature and better rule. FOIL+Gini means using information gain and gini coefficient to do feature select and rule select. FOIL+Gini+Support means using chapter 3.2 research method to improve our rule. And the data are the number of the transactions. The comparison of the confidence, support and coverage about the different method are shown in Figures 1 to 3.

![Figure 1. Confidence Comparison](image-url)
From the experimental results it can be seen that about the confidence of the rules, the confidence of the method combining the information gain, support and Gini coefficient is no higher than the other method and Gini coefficient has almost no effect on experimental result. However, in the current situation of internet finance, the internet financial anti-fraud rules need higher support. The origin rules’ coverage and support is high, but the confidence is too low due to its generality. By combining the FOIL information gain, Gini coefficient and support degree, about the rule coverage and support this paper’s method is better than the other method, and maintains a relative high confidence level at the same time. This method has good performance in support. For the background of the internet financial anti-fraud, this paper’s method is some effective.

5. Conclusion
In the current situation of internet financial anti-fraud, it is difficult to choose a better attribute in the process of improving the fraud rules by single metric. And the traditional method of evaluating the effectiveness of the rules cannot adapt to the internet financial anti-fraud demand. This paper aims at the demand of bank internet financial anti-fraud system and fraud detection, combines the FOIL information gain, support degree and Gini coefficient to improve the rules in the internet financial anti-fraud field. The experimental results show that the proposed method has good performance in support, and has some improvement for the internet financial anti-fraud.
6. References

[1] Fu K, Cheng D, Tu Y and Zhang L 2016 Credit Card Fraud Detection Using Convolutional Neural Networks International Conference on Neural Information Processing pp 483- 490 Springer Cham 5

[2] Bahnsen A C, Aouada D and Stojanovic A 2016 Feature engineering strategies for credit card fraud detection Expert Systems with Applications An International Journal 51(C) 134-142

[3] Sahin Y, Bulkan S and Duman E 2013 A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications 40 (15) 5916-5923

[4] Bahnsen A. C, Stojanovic A, Aouada D, and Ottersten B 2014 Cost Sensitive Credit Card Fraud Detection Using Bayes Minimum Risk. International Conference on Machine Learning and Applications Vol.1 333-338 IEEE

[5] Bhingarde A, Bangar A, Gupta K and Karambe S 2015 Credit card fraud detection using hidden markov model. International Journal of Computer Science & Information Technology 76(3) 169-170

[6] Quinlan J R and Cameron-Jones R M 1993 FOIL A midterm report Machine Learning ECML-93 Springer Berlin Heidelberg

[7] Imielienskin T, Swami A and Agrawal R 1993 Mining association rules between set of items in largedatabases Acm Sigmod Record 22(2)

[8] Quinlan J R 1993 C4.5: programs for machine learning 1

[9] Srikant R. and Agrawal R 1995 Mining Generalized Association Rules International Conference on Very Large Data Bases Vol.13 407-419

[10] Breiman L, Friedman JH, Olshen RA and Stone CJ 1984 Classification and regression trees (cart) Encyclopedia of Ecology 40(3) 582-588