Anomaly Detection Algorithm Based on Semi-Supervised Collaborative Strategy

Mingtian Shao\textsuperscript{a} and Naijie Gu\textsuperscript{b,}\textsuperscript{*}

School of Computer Science and Technology, University of Science and Technology of China, Hefei, China

E-mail: \textsuperscript{a}mitty@mail.ustc.edu.cn, \textsuperscript{b,}\textsuperscript{*}gunj@ustc.edu.cn

Abstract. This paper analyzes the existing credit card anomaly detection classification algorithms, summarizes the parts that can be improved, and proposes an outlier detection classification algorithm based on the unsupervised algorithm and active learning decision trees. Since the 1990s, machine learning technology has been widely used in the field of credit card fraud detection. Among them, the supervised learning expert system used for classification tends to perform better than the unsupervised learning model. The good performance of supervised learning methods requires high learning costs. The training process is usually serial or partially parallel. Therefore, the computing power, time, and manual labeling cost for learning need to be considered as algorithm selection factors. However, manual labeling usually involves the counter or ATM sending information to the back-end large-value transaction authorization department for manual review, and then the employee judges whether it is an abnormal transaction based on experience. This rigorous process will produce a very large time delay. Combining this feature, this article selects the appropriate unsupervised outlier detection method, selects part of the training data, and uses a small part of the more valuable data for annotation learning. Experiments show that this method can improve the accuracy while saving the time cost of training. Compared with the training time, the added time of classification is negligible.

1. Introduction

In order to improve the recognition accuracy of anomaly detection classifiers, a multi-level classification model based on a semi-supervised collaborative strategy is designed and implemented. This method uses the unsupervised algorithm of covariance estimation as the first-class classifier to roughly classify the experimental data. The improved decision tree algorithm based on migration learning is used as the secondary classifier to realize the classification model that reduces the cost of labeling and improves the recognition accuracy. By adjusting the two custom model configuration parameters set: the primary classification division parameter and the second classification migration parameter, the results on multiple data sets show that the classification effect has been improved. The improved method performs better than other algorithms on data sets such as RWC. Taking the RWC data set as an example, compared to the XGBoost algorithm AUC, there is an increase of 1.88%, F1-score has an increase of 1.12%, and the overfitting ratio also drops from 0.1433 to 0.1071.


2. Related Work

There are many ways to solve the problem of abnormal credit card transaction classification, which can be divided into two types: single method and combined method. The recognition methods of a single model mainly include Bayesian, artificial BP network, Decision Tree (DT), clustering, recognition based on association rules, and Support Vector Machines (SVM). The combined learning algorithm is an integration of a single classification method, and the combined methods include Bagging [1], Boosting [2], etc. This chapter will introduce a series of single model methods suitable for abnormal credit card transaction detection, as well as several examples of integrated methods [3].

2.1. Single model classification algorithm

Different data sets have completely different feature distributions, depending on the characteristics of the target variable [4], resulting in a differentiated single classification module [5]. DT’s training process parameter adjustment is based on sample-based summary and induction [6]. This method is a top-down iterative method. It performs feature comparison and classification at each bottom-level branch point [7]. SVM is a classification method based on the related ideas of inductive statistics [8]. The main structure of NN is to simulate the connection combination similar to the synapse of human brain neuron [4], and it is a kind of mathematical algorithm for combined data analysis. A Bayesian model such as Naive Bayes is an algorithm that uses probability theory and statistical knowledge to classify [2].

2.2. Integrated model classification algorithm

In practical applications, the complexity and diversity of data structures make a single classification method often unable to achieve ideal results. Therefore [7], the fusion of multiple classification methods has become a breakthrough [9], that is, the integrated learning method. The integrated method is regarded as one of the four main machine learning development trends at present and has become a research hotspot in a wide range of international scientific research and business circles [10]. There are rich methods for combining several independent learners, such as weighted voting. Common algorithms include Bagging, Boosting, etc., among which Boosting is mainly serial to reduce deviation, while Bagging method is To reduce variance in parallel, the two have an opposite relationship.

3. Algorithm design and implementation

This training method To some extent it is similar to semi-supervised learning [11]. The structure of the two-level classifier is shown in Figure 1. Since the abnormal data output by the first-level classifier is actually abnormal, the output is normal and there are more data that are actually normal. Therefore, the second-level division module requires targeted adjustments when dividing different input data. Divide the module parameters.

The first-level division module is denoted as $f_1$, and the second-level classifier is denoted as $f_2$. The proportion of the number of samples identified as abnormal by $f_1$ in the training data $S'$ to the total number of samples in $S'$ is $\theta$, the number must be greater than the number of actual fraud data points in the total sample set. It can be expressed by a formula as (1). In addition, the composite semi-supervised collaborative classifier can be expressed as $F_{ensemble} = f_2(f_1)$.

The training of $F_{ensemble}$ is mainly the adjustment of the parameter $\theta$ and the supervised training of $f_2$.

$$\theta = \frac{S'(f_1(S') = 1)}{S'} > \frac{S'_-}{S'}$$

(1)
3.1. Outlier detection algorithm
Detecting outliers is an extremely valuable process. Because in this process, additional information about the distribution of the data set can be obtained, so that the data can be roughly processed [12]. For outlier detection, the more accurate the recognition results of these methods, the better [12], but in this article, the common ensemble algorithm: Robust Covariance (RC) [13] is used as first-level rough classifiers. Therefore, the evaluation and selection of recognition models must be based on other new standards. Considering the impact on the final detection results, the smaller the range of outliers in the first-level classification results, the better, and the true outliers should be included as much as possible. The formula can be expressed as (2). The following brackets are preconditions. Compared with normal recognition, the ratio of the number of data points recognized as abnormal to the total number of data points can be appropriately relaxed $\theta$. This ratio can be manually adjusted to achieve the best detection effect. At this time, the formula can be updated to (3).

$$\min \frac{\text{Number of data points predicted to be abnormal}}{\text{Total data points}} \quad (\max \frac{\text{Number of data points predicted to be abnormal}}{\text{Actually anomalous samples}})$$

(2)

$$\max \frac{\text{Number of abnormal samples predicted}}{\text{Actual number of abnormal samples}} \quad \left(\frac{\text{Number of abnormal samples predicted}}{\text{Number of samples in the entire data set}} = \theta\right)$$

(3)

3.2. Two-level classification based on decision tree
The secondary classifier uses the modified cost-sensitive decision tree algorithm, which can well solve the problem of large differences in classification results when a single sub-classifier faces data sets with different data distributions. This strategy is based on the idea of transfer learning. When the approximate range of the data distribution is known, the output results are controlled by adjusting and setting custom pruning parameters and regression partitioning parameters. Set the regression partition parameter to $\epsilon$. There are several important reasons for choosing a decision tree as a secondary classifier: the decision tree can easily integrate the weights of data points into the model learning process, and it is not necessary to adjust the weights of data points by sampling. And the decision tree has better expression and generalization possibilities, which can be compromised by adjusting the number of layers of the tree.
In fact, if you do not consider the labeling and learning costs, you can also use the entire data set to train to obtain a true fuzzy range, and combine the false fuzzy range of the secondary training set to adjust the classifier to obtain more accurate results. Add the adjustable parameter $\epsilon$ to the divided segmentation value vector as the adjustment parameter of transfer learning. Here we improve the regression tree and set $\epsilon$ to 1 when training on the training set. Assuming that the data set has $M$ samples and $N$ attributes, where $\text{ave}_i$ represents the average value of the training set elements in the leaf, and the input space divided into two is represented as $R_1(c, d \cdot \epsilon) = x \mid x^c \leq d \cdot \epsilon$, $R_2(c, d \cdot \epsilon) = x \mid x^c > d \cdot \epsilon$.

4. Experimental platform and data
Before conducting experiments to analyze the experimental results, this section introduces the experimental environment and data sets. Following the principle of controlling variables, the same experimental platform and experimental data are used for comparative experiments of different algorithms during the experiment.

4.1. Experiment platform
The experiment is carried out on the local machine, and the parameters of the experiment platform are shown in the table 1.

| Table 1. Classifier optimization experiment platform |
|-----------------------------------------------|
| **GPU Server with 8 Graphics Cards**          |
| CPU                                           |
| 2 Intel XeonE5-2620 v2 2.1GHZ CPUs           |
| GPU card                                      |
| NVIDIA Tesla K40m                            |
| Single-precision peak                         |
| 4.29 TFLOPS                                   |

4.2. Experimental data
The data used in the experiment is shown in the table 2, and the method of 5 divisions and cross-validation is also used, 80% of the samples are training samples, and the remaining 20% are validation samples. The experimental samples undergo uniform sample preprocessing, encoding, outlier elimination and normalization. The P:N means the Ratio of Positive:Negative.

| Table 2. Dataset used in the experiment |
|----------------------------------------|
| **Dataset**                           |
| **Sample Size**                       |
| **N:P(1:0)**                          |
| **Data Dimension**                    |
| Real World Credit (RWC)               | 284807 | 492 : 284315 $\approx$ 1 : 577.88 | 31 |
| Kaggle Credit card (KC)               | 3075  | 448 : 2627 $\approx$ 1 : 5.86      | 9  |
| UCI Statlog German Credit (USGC)      | 1000  | 300 : 700 $\approx$ 1 : 2.33       | 61 |

5. Experiment and result analysis
In order to test the effect of the anomaly detection classifier proposed in this article, first observe the effect of parameter values on the experimental results under the condition of setting different custom migration parameters on the RWC data set. Then compare with the unimproved XGBoost classifier and other decision tree algorithms and classic classification algorithms to verify the effectiveness of the improved strategy. Finally, repeat the experiment on the other two sample sets to test the universality of the model.
5.1. Evaluation index

The first-level classifier uses the recall rate as the main evaluation index, and the second-level classifier experiment also uses AUC and F1-score as the evaluation index, which is the same as the definition of each evaluation index given in the other credit card data classification paper [6]. Finally, for the total model composed of the two-level classifier, the three indicators of calculation time, AUC and F1-score of the classification on the test set are evaluated. Then, we set another evaluation index to judge the overfitting of the model, as shown in Equation (4).

\[
\text{Fit Ratio} = \left| \frac{\text{TrainAverageAUC}}{\text{TestAverageAUC}} - 1 \right| \quad (4)
\]

5.2. Model test results

After the parameter tuning and pre-training of all the classification modules, the comparison experiments between the method in this paper and other basic tree methods and benchmark methods are carried out on the test set of data sets such as RWC. The experiment selection is compared with six methods GBDT, XGBoost, CART, RF, LOR and SVM. Among them, CART and RF are decision tree models for comparison. GBDT and XGBoost are both integrated models based on their fundamental optimization. LOR and SVM is a more classic machine learning algorithm.

| Table 3. Comparison of the results of multiple classification methods in the RWC data set |
|---------------------------------|-------|--------|--------|-------|-------|-------|-------|
|                                | EI    | Classifier | Ours   | XGBoost | GBDT  | CART  | RF    | SVM   | LOR   |
| AUC                            | 0.8895| 0.8707   | 0.86   | 0.7417 | 0.8569| 0.8195| 0.78  |
| F1-score                       | 0.8725| 0.8613   | 0.8182 | 0.7322 | 0.8507| 0.5333| 0.6942|
| Fit Ratio                      | 0.1071| 0.1433   | 0.1623 | 0.1131 | 0.2285| 0.2185| 0.2813|

As shown in the table 3, the recognition results of the model in this paper are shown in italics. EI means Evaluation Index. From the table, we can see that the results of the improved method on the RWC sample set are better than other algorithms. Compared with the GBDT algorithm, the AUC has an increase of 1.88%, the F1-score has an increase of 1.12%, and the overfitting ratio has also dropped from 0.1433 to 0.1071. And by outputting the intermediate results of the first-level classification on the test set, it can be found that most abnormal transactions are classified as abnormal samples in the first-level classification, which is in line with the expected high recall rate of abnormal samples. The actual recall rate is 93.17%. The Figure 2 is the dimensionality reduction output of the first-level classification result and the original test set category. From the picture, the effect of the first-level classification can be more intuitively felt.

![Figure 2. First-level division test results](image-url)
Table 4. Comparing experimental results with other data sets of the model

| Classifier | Dataset | EI | Ours | XGBoost | GBDT | CART | RF | SVM | LOR |
|------------|---------|----|------|---------|------|------|----|-----|-----|
|            | KC      | AUC | 0.9322 | 0.9112 | 0.8820 | 0.8675 | 0.8255 | 0.8714 | 0.8456 |
|            |         | F1-score | 0.8714 | 0.8602 | 0.8534 | 0.8162 | 0.8315 | 0.8503 | 0.8431 |
|            |         | Fit Ratio | 0.0424 | 0.0518 | 0.0785 | 0.1467 | 0.2123 | 0.1332 | 0.1887 |
|            | USGC    | AUC | 0.73 | 0.7182 | 0.6740 | 0.6953 | 0.6600 | 0.6523 | 0.6667 |
|            |         | F1-score | 0.5893 | 0.5771 | 0.5405 | 0.5320 | 0.5400 | 0.5213 | 0.5235 |
|            |         | Fit Ratio | 0.0712 | 0.0800 | 0.1053 | 0.1093 | 0.1200 | 0.1002 | 0.1792 |

To further verify the effectiveness of the model, the results on the KC and USGC sample sets are shown in table 4. The accuracy on the KC data set is increased to 93.22%, and that on the USGC data set is increased to 73%, which is 2.1% and 1.18% respectively compared with the XGBoost benchmark method.

6. Conclusion

The method introduced in this paper is a secondary classification model based on a semi-supervised collaborative strategy. The unsupervised classification method is selected as the primary classifier, and the improved decision tree method based on active learning and migration learning is implemented as the secondary classifier. The idea of a semi-supervised algorithm can effectively save the time of manual labeling, which is helpful to the practical application of machine learning methods. Besides, the choice of the first-level classifier in this article is only an experiment on some classic outlier detection algorithms, and it is not ruled out that there are algorithms that are more suitable for the model in this article. This paper proposes that the model only proves the effectiveness of the secondary classification structure, and the details of the model still have room for optimization.

7. References

[1] Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke K. Nandi. Credit Card Fraud Detection Using AdaBoost and Majority Voting. 6:14277–14284.
[2] Yufei Xia, Chunzhe Liu, YuYing Li, and Nana Liu. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. 78:225–241.
[3] Anuj Sharma and Prabin Kumar Panigrahi. A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. 39(1).
[4] David West. Neural network credit scoring models. 27(11):1131–1152.
[5] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudik, editors, AISTATS, volume 15 of JMLR Proceedings, pages 315–323. JMLR.org.
[6] Gang Wang, Jian Ma, Lihua Huang, and Kaiquan Xu. Two credit scoring models based on dual strategy ensemble trees. 26:61–68.
[7] Tian-Shyug Lee, Chih-Chou Chiu, Yu-Chao Chou, and Chi-Jie Lu. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. 50(4):1113–1130.
[8] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. 33(4):847–856.
[9] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. 42(4):463–484.
[10] Mark D. Schluchter and Janet T. Elashoff. Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. 37(1-2):69–87.
[11] Xiaojin Zhu and Andrew B. Goldberg. Introduction to Semi-Supervised Learning. 3(1):130.
[12] Peer Kroger. Outlier Detection Techniques.
[13] Peter, J, and Rousseeuw. Least Median of Squares Regression.