Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 10, No. 1, 2023

Research on Financial Fraud Identification System Based on Differential Privacy

Sijia Shen *, Yueling Li

College of Science, China University of Petroleum (East China), Qingdao, Shandong, 266580, China

* **Corresponding author**: Sijia Shen

Abstract: Data sharing among financial institutions is often not possible, resulting in a "data silo" situation. In this paper, we train an efficient financial fraud detection model under the framework of privacy protection from the perspective of facilitating the collaboration of multiple financial institutions to train a fraud identification system. This paper first used traditional oversampling and under sampling methods to balance the data and train models such as logistic regression, support vector machines and random forests, but did not obtain the desired results. In contrast, the optimal subsampling method based on logistic regression performs well in terms of training results and program runtime. To protect data security, differential privacy is introduced on this basis to find the classification accuracy of the model under different privacy budgets. It is concluded that a certain balance between the degree of privacy protection and model effectiveness should be achieved according to privacy requirements.

Keywords: Fraud Detection; Isolated Data Island; Optimal Subsampling; Differential Privacy.

1. Introduction

With the booming development of digital finance in recent years, the ever-changing fraudulent methods have placed higher demands on the anti-fraud capabilities of financial institutions, becoming a core capability that all types of financial institutions must build in the digital financial era.

The use of large-scale financial data is bound to encounter the following problems: firstly, ensuring data security and user privacy are not compromised; secondly, privacy protection, "data silos" and other phenomena that prevent the sharing of data. Due to the lack of data sharing, good predictions and early warnings cannot be obtained using a single dataset to meet the high demands of financial institutions' anti-fraud efforts. There is an urgent need to collaborate with many financial institutions to build an efficient and robust financial fraud identification system, while protecting user privacy and complying with privacy regulations.

2. Method

2.1. Unbalanced data processing

The imbalanced dataset has an uneven distribution of label classes. The part of the sample with a low proportion of labels provides less information, resulting in poor classification of the model. Therefore, some degree of treatment of the imbalanced data is required, which is usually done from the perspective of the data.

2.1.1. Classical sampling methods

(1) Oversampling

For unbalanced data, one idea is to repeatedly sample a small number of labelled samples. The Synthetic Minority Oversampling Technique (SMOTE) algorithm artificially synthesises new minority samples in a certain way. The principle is that for each sample point x in the minority class, find its k nearest neighbour samples, and select N samples at random among them, and perform random linear interpolation between the sample point x and these N nearest neighbours to construct a new minority class sample. However, repeated sampling of minority class samples to generate new samples increases the possibility of overlap between classes and is prone to the problem of sample overlapping (Overlapping), which may generate samples that do not provide useful information [1-3].

(2) Undersampling

Another approach is to reduce the majority class samples to balance the data. Undersampling does not expand the original training set, does not lead to an increase in computation, and does not result in sample overlap, which solves the problems that exist when using oversampling for large data sets. However, the samples that are removed using this method may contain important information and result in information loss.

2.1.2. Optimal subsampling for Large Sample Logistic Regression

Since computation is a bottleneck for the application of logistic regression on massive data, Wang H et al. (2017) proposed a method for optimal subsampling. According to the general subsampling algorithm, a more efficient subsampling procedures is proposed by choosing nonuniform $\pi_i's$ to "minimize" the asymptotic variance-covariance matrix.

• Minimum Asymptotic MSE of $\tilde{\beta}$

The asymptotic MSE of $\tilde{\beta}$ is equal to the trace of V, namely,

$\mathrm{AMSE}(\tilde{\beta}) = \mathrm{tr(V)}$.

Denote $V = M_x^{-1} V_c M_x^{-1} = O_P(r^{-1})$, $V_c = \frac{1}{rn^2}\sum_{i=1}^{n} \frac{\{y_i - p_i(\hat{\beta}_{MLE})\}^2 x_i x_i^T}{\pi_i}, M_x = n^{-1}\sum_{i=1}^{n} w_i(\hat{\beta}_{MLE})\, x_i x_i^T$. V depends on $\{\pi_i\}_{i=1}^{n}$. The key idea of optimal subsampling is to choose nonuniform subsampling probability (SSP). Since minimizing the trace of the (asymptotic) variance-covariance matrix is called the A-optimality criterion, the resultant SSP is A-optimal in the language of optimal design.

If the SSP is chosen such that

$$\pi_i^{\mathrm{mMSE}} = \frac{|y_i - p_i(\hat{\beta}_{\mathrm{MLE}})| \|\mathrm{M}_X^{-1}\mathrm{x}_i\|}{\sum_{j=1}^{n} |y_j - p_j(\hat{\beta}_{\mathrm{MLE}})| \|\mathrm{M}_X^{-1}\mathrm{x}_j\|}, i = 1,2,,n,$$

then the asymptotic MSE of $\tilde{\beta}$, tr(V), attains its minimum.

• Minimum Asymptotic MSE of $M_X\tilde{\beta}$

Instead of focusing on the more complicated matrix V, an alternative optimality criterion by focusing on V_c is defined according to the partial ordering of positive definite matrices. That is to minimize $\mathrm{tr}(V_c)$, instead of minimizing $\mathrm{tr}(V)$.

If the SSP is chosen such that

$$\pi_i^{mVc} = \frac{|y_i - p_i(\hat{\beta}_{MLE})|\|x_i\|}{\sum_{j=1}^{n}|y_j - p_j(\hat{\beta}_{MLE})|\|x_j\|}, i = 1,2,\dots,n,$$

then $\mathrm{tr}(V_c)$ attains its minimum.

2.2. Differential Privacy

Differential Privacy (DP) is an approach that promises to combine privacy protection with the desire to gain maximum insight. The main idea behind differential privacy is to randomly perturb the results of an analysis so that any individual in the data has a negligible impact on the results. The concept of differential privacy, developed by Dwork et al. (2012) at Microsoft, uses random noise (random perturbations around the true value) to construct data that makes personal information difficult to identify while retaining statistical properties.

An important component of differential privacy calculations is the privacy loss parameter, usually denoted by ε, which determines the amount of noise added to the calculation; the larger the ε, the more accurate the statistical result but the less privacy preserving it is.

A random function K is said to satisfy (ε,δ)-differential privacy if the probabilistic output of the function K on some adjacent data set D1 and D2 satisfies the following inequality.

$$\Pr[K(D_1) \in S] \leq \exp(\varepsilon)\Pr[K(D_2) \in S] + \delta$$

Differential privacy algorithms generate noise from a Laplace distribution. ε directly controls the variance of the distribution, and thus its randomness. A common approach to adding noise to differential privacy algorithms is to use the Laplace mechanism, mainly for numerical data. For any given query function $f: D \to R^d$, if M(D) satisfies

$$M(D) = f(D) + (Laplace\left(\frac{\Delta f}{\varepsilon}\right))^d$$

then the Laplace mechanism satisfies ε-differential privacy. ε directly controls the variance of the distribution and thus its randomness [7-8].

3. Research and Analysis

3.1. Data Sources

This paper uses a comprehensive financial dataset generated by the PaySim mobile money simulator, a dataset specifically designed for fraud detection research. The dataset has a total of 6,362,620 data and 11 variables, the specific variable names and their descriptions are shown in the table below. The variable "is Fraud" is the target variable of interest, indicating whether it is fraudulent or not.

The dataset used in this paper is extremely unbalanced data with 99.87% of the not Fraud samples and less than 0.13% of the Fraud samples. For extremely unbalanced data, direct model training may not yield satisfactory results.

3.2. Model Study

3.2.1. Financial fraud classification prediction based on traditional sampling methods

If the data is not processed for the imbalance nature, the classification accuracy of each model is not satisfactory after the data is normalized and the test and training sets are divided, indicating that the models obtained from the unprocessed data training may have the risk of overfitting. The results of the four models took a total of 5182.9347 seconds to obtain.

The SMOTE method was used on the training set to expand the samples with few labels in the original training set to obtain a new training set with a relatively balanced ratio of two types of samples, and multiple classifiers were used on the new training set for comparison. The accuracy of the models obtained using the oversampled data was significantly improved compared to the models trained on the unsampled data, especially for logistic regression, Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT). Since the idea of oversampling is to repeatedly sample a small number of labelled samples, the two types of data become balanced while making the originally large-scale data even larger, resulting in a significant increase in program running time.

Random under sampling is applied to the same processed training set, which achieves compression of the original training set by sampling most classes of samples to obtain a relatively balanced new training set. Again, the performance of models trained on under sampled data has improved significantly compared to those trained on unsampled data, with SVM, random forest and GBDT performing extremely well, with accuracy rates approaching 1. While both sampling methods improved over the unsampled models, under sampling improved more. Although random under sampling performed well and was much faster to compute, taking only 786.3378 seconds, but the improvement in run speed was obtained by losing the integrity of the data.

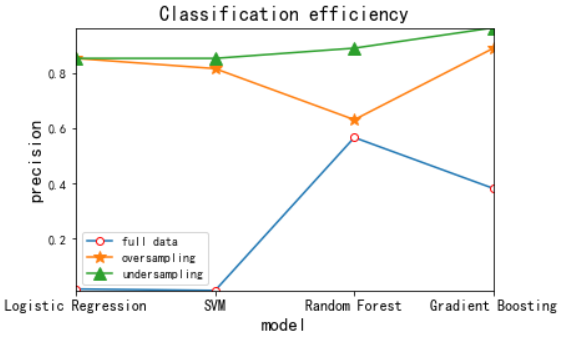
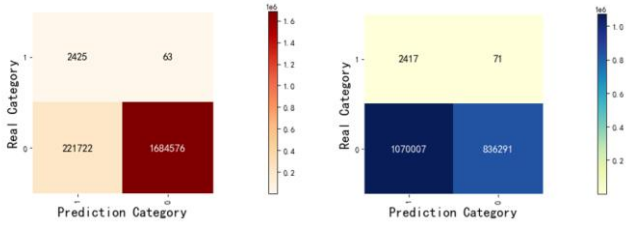


Fig. 1 Classification accuracy of each model using traditional sampling methods

3.2.2. Financial fraud classification prediction based on optimal subsampling method

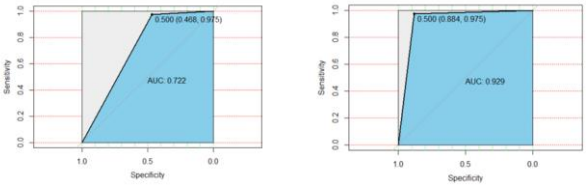
Here the imbalanced nature of the financial fraud dataset is treated using optimal subsampling method, and then a logistic regression model is used to complete the classification. Uniform subsampling is used for comparison. The classification confusion matrix of the sampled model is calculated separately.



(a) Optimal subsampling (b) Uniform subsampling

Fig. 2 Confusion matrix for classification using different sampling methods

By looking at the confusion matrix it was found that some of the non-fraudulent high-risk behaviors are mistakenly flagged as fraudulent in both classification results, but this early warning of risk is also essential. The optimal subsampling method performs significantly better than the uniform subsampling method.



(a) Optimal subsampling (b) Uniform subsampling

Fig. 3 ROC curve for classification using different sampling methods

A major advantage of the optimal subsampling method is that it uses Newton's method of optimization, which has relatively low computational complexity. Using logistic regression-based optimal subsampling can overcome the computational bottlenecks caused by the explosive growth in data volume, and optimal subsampling takes significantly less time to run than traditional sampling methods to train the model and obtain the classification results without loss of accuracy. Here the optimal subsampling method runs the program in 46.1970 seconds, which is only slightly longer than the 27.1651 seconds of the uniform subsampling method, but has better model performance than uniform subsampling.

As the empirical results show, the optimal subsampling method for logistic regression is a computationally feasible method for very large samples, and the sampling results are excellent, yielding excellent classification models that approximate the results based on all data well.

3.2.3. Financial Fraud Identification Systems in a Differential Privacy Framework

In this section, the data set is randomly broken up and divided into three financial institutions (banks are used as an example). 70% of the data set is selected as the training set and the remaining 30% is the test set. The evaluation metric used for the test results in this section is accuracy, which is the proportion of the total number of data where the predicted and actual values are the same to the total number of data.

A differential privacy mechanism is used to incorporate Laplace noise to protect banks' privacy. Separate experiments are conducted on three data sets to analyse the impact of different privacy budgets ε on the prediction results and to explore the performance of the differential privacy model. In the Laplace mechanism, the noise is randomly generated by the probability density function, and the experimental accuracy obtained after adding noise is also random in nature. Therefore, this section finds their average accuracy under different privacy budgets.

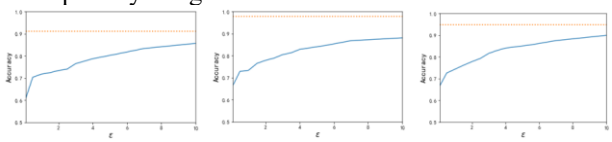


Fig. 4 The impact of privacy budgets ε on accuracy on the 3 datasets

The introduction of differential privacy reduces the effectiveness of prediction compared to the accuracy of ordinary logistic regression models, and the smaller the privacy budget, the lower the prediction accuracy, which is the cost of performing privacy preservation. At the same time, the larger the privacy budget, the less noise is added, the greater the classification accuracy of the model and the higher the utility of the predictions. This suggests that protecting an institution's privacy requires sacrificing a certain level of accuracy. Therefore, a balance should be struck between the performance of the model and the degree of privacy protection, depending on the privacy needs required by the institution.

4. Conclusions

Existing fraud identification systems use internal data collected separately by each financial institution and bank, which cannot be shared between institutions due to privacy protection and "data silos". Secondly, financial transaction data is extremely unbalanced, with the number of fraudulent transaction samples far less than the number of normal transaction samples, which to a certain extent affects the predictive effect of financial fraud models. In order to address these problems, this paper uses differential privacy and optimal subsampling to solve them, and proposes a framework for financial fraud identification system under the differential privacy framework for real-life financial fraud scenarios. The conclusions of this paper can be summarized in the following two parts.

(1) The optimal subsampling method is proposed to be applied to large-scale imbalanced data processing. When the data volume is large and the proportion of positive and negative samples is disparate, the traditional methods of oversampling and under sampling sometimes fail to achieve good results. Experimental results show that the optimal subsampling method for logistic regression is a computationally feasible method for very large samples and can approximate the results based on all data well.

(2) A framework for a differential privacy-based financial fraud identification system is proposed, which provides a new idea for solving the "data silo" problem. The experimental results show that the classification accuracy of the model decreases to a certain extent when compared with the prediction results of logistic regression without noise interference, and the utility of the model increases with the increase of privacy budget. This requires a balance between the degree of privacy protection and the effectiveness of the model classification according to privacy requirements.

References

[1] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Oversampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.

[2] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017,73:220–239.

[3] Elreedy D, Atiya A F. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance - ScienceDirect[J]. Information Sciences, 2019, 505:32-64.

[4] Wang H, Zhu R, Ping M. Optimal Subsampling for Large Sample Logistic Regression[J]. Journal of the American Statistical Association, 2017, 113(2):1440037-1438957.

[5] Dwork C. Calibrating noise to sensitivity in private data analysis[J]. Lecture Notes in Computer Science,2012,3876(8):265-284.

[6] Zhang J , Zhang Z , Xiao X , et al. Functional Mechanism: Re gression Analysis under Differential Privacy[J]. Proceedings o f the VLDB Endowment, 2012, 5(11):1364-1375.

[7] Kurz Christoph. Understanding differential privacy[J]. Signifi cance, 2021, 18(3):24-27.

[8] Samet S. Privacy-preserving logistic regression[J]. Journal of Advances in Information Technology,2015,6(3):1–8.

[9] Alonso E, Elmir A, Axelsson S. Paysim: a financial mobile money simulator for fraud detection[C]// European Modeling & Simulation Symposium. 2016.