Credit card fraud detection based on machine learning

Tran Thanh Cong*
Hong Bang International University

ABSTRACT
Online transactions have increased dramatically over the decades. Credit card transactions account for a large proportion of these transactions. This leads to an increase in credit card fraud transactions, causing damage to the financial industry. Therefore, it is important to create fraud detection systems, consisting of two labels fraud and no fraud. However, the dataset is not balanced between the two labels. In this paper, we use the resampling method such as SMOTE to process this unbalanced dataset to obtain a balanced dataset. The machine learning (ML) algorithms, named random forest, k nearest neighbors, decision tree, and logistic regression are applied to this balanced dataset to create ML models. The performance of these ML models is evaluated through accuracy, recall, precision, and F1 score. We observed that the SMOTE-based random forest algorithm identifies frauds in a better way than other algorithms.

Keywords: *machine learning, classification, imbalanced data, SMOTE, fraud detection*

1. INTRODUCTION

Fraudulent activities have been increasing significantly in various industries worldwide, particularly in the financial industry. In financial companies, credit card fraud is considered the most problematic and is needed to prevent it as soon as possible. In order to reduce dramatically consequences of credit card fraud, fraud detection approaches need to investigate to strictly handle. Systems of fraud detection are trained through older transactions so as to decide about future ones [1].

In fraud detection, the number of normal cases is significantly more than fraudulent cases. This leads to the status of imbalanced data. In the skewed dataset, one class of data has a very high number of instances while the other class accounts for a very small number of ones. However, machine learning algorithms work effectively on the balanced distribution of classes. In order to tackle the issue of the skewed datasets, many solutions have been researched in the past few years. In these researches, three groups, known as data-level, algorithm-level, and ensemble solutions, are commonly proposed [2].

In this paper, the resampling approach, named Synthetic Minority Over-sampling Technique (SMOTE) is used to obtain a balanced dataset. The machine learning algorithms, named random forest (RF), k nearest neighbors (KNN), decision tree (DT), and logistic regression (LR), then, are utilized to train the balanced dataset obtained. Comparisons of performances of machine learning algorithms based on two resampling techniques are pointed out to select the best case to detect credit fraud.

The remainder of this paper is structured as follows: Section 2 presents related work. Section 3 is about the description of dataset. Section 4 shows SMOTE technique. Section 5 describes machine learning algorithms and classification measurements. Section 6 illustrates the outcomes. Lastly, section 7 concludes this paper.

2. RELATED WORKS

Credit card fraud has resulted in a huge loss in both customers and financial companies worldwide. Therefore, researchers have an effort to search for optimized methods to detect and prevent this fraud. Recently, machine

Corresponding author: M.S. Tran Thanh Cong
Email: congtt@hiu.vn

learning approaches have applied to detect fraudulent activities. In the paper [3], the approach named extreme outlier elimination and hybrid sampling technique is proposed to generate high predictions for both fraud and non-fraud classes. In the paper [4], P. Kumari et al. analyzed several ensemble classifiers, known as Bagging, Random Forest, Classification via Regression, Voting and compared them with some effective single classifiers such as K-NN, Naïve Bayes, SVM, RBF Classifier, MLP, Decision Tree. The evaluation of these algorithms is implemented through three different datasets and treated with SMOTE to solve the skewed dataset. In the [1], D. S. Sisodia et al. indicated that the SMOTE ENN method detects the fraud in a better way than other classifiers in the set of oversampling techniques considered, and TL works better on the set of undersampling techniques taken. In the paper [5], R. Brause et al. proposed the combined probabilistic and neuro-adaptive method for a given database of credit card transactions of the GZS. In the paper [6], A. Srivastava et al. used a hidden Markov model (HMM) to model the sequence of operations in credit card transaction processing and showed how it can be used for the detection of frauds. In the paper [7], S. Benson Edwin Raj et al. analyzed several modern techniques based on Artificial Intelligence, Data mining, Fuzzy logic, Machine learning, Sequence Alignment, Genetic Programming to detect various credit card fraudulent transactions. These approaches proved efficient credit card fraud detection system.

In summary, there are a variety of ways to create models for identifying fraudulent activities in order to reduce risks in the financial field. However, depending on each specific case, these techniques mentioned in this paper have different strengths and weaknesses. In this paper, we use four simple but effective ML algorithms to predict the fraud activities. In addition, apart from accuracy measurement for ML models, we also investigate more measures such as recall, precision, and F1 score to provide better evaluation measures.

3. DATASET

In this study, we select dataset of credit card fraud prediction as a case study of imbalanced dataset obtained in [8]. This dataset included transactions which is collected within 2 days and created by credit cards in September 2013 by European cardholders. The dataset has 284,807 transactions and 31 columns. These columns consist of 28 numerical features, named V1 to V28, a feature of "Time", a feature of "Amount" and a label of "Class". The label of "Class" is divided into two classes, such as positive class (frauds) denoted 1, and negative class (no-frauds) denoted 0. This dataset is highly skewed. The positive class (frauds) occupied 0.172% of all transactions which is revealed in **Figure 1.**

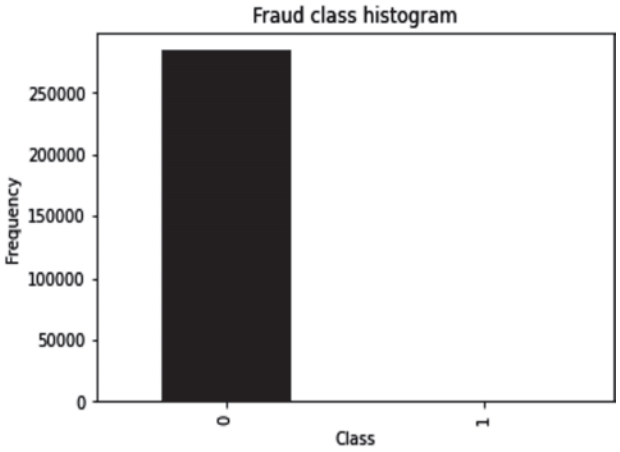

Figure 1. *Fraud class histogram*

After analyzing the data set through python 3.7 version, we recognized that the feature "Amount" has values from 0 to 25,691.16. However, if features are on a relatively similar scale and/or close to normally distributed, the ML algorithms will have good performances or converge faster. Therefore, standardization is used to eliminate the mean and scale to unit variance to alleviate the wide range of the feature "Amount", so approximately 68% of the values lie in between (-1, 1) [9]. A new feature, named "normAmount" is added to the dataset. Additionally, the "Time" and "Amount" features do not need to build model so we will drop these features. We obtained the dataset of 30 features and the "Class" label.

Next, in order to deal with the issue of imbalanced data, the resampling technique, named synthetic minority oversampling technique (SMOTE) [10] is used in this study. As the SMOTE techniques is popular and has demonstrated their benefits when solving skewed dataset [11 - 12]. After implementing these techniques for the credit fraud detection dataset based on python 3.7 version, we accomplished the balanced dataset.

4. SMOTE TECHNIQUE

SMOTE technique which is one of popular oversampling techniques can mitigate the risk of overfitting faced by random oversampling [13]. This technique produces artificial samples instead of merely duplicating current observations. As illustrated in **Figure 2**, this is obtained through linearly interpolating a randomly selected minority observation and one of its neighboring minority observations. In other words, this technique has three steps to produce a synthetic sample. The first step is to select a random minority observation $\vec{a}$. The second step is to select instance $\vec{b}$ in its *k* nearest minority class neighbors. The last step is to create a new sample $\vec{x}$ through randomly interpolating the two samples: $\vec{x} = \vec{a} + w \times \left(\vec{b} - \vec{a}\right)$, where w is denoted as a random weight in [0,1].

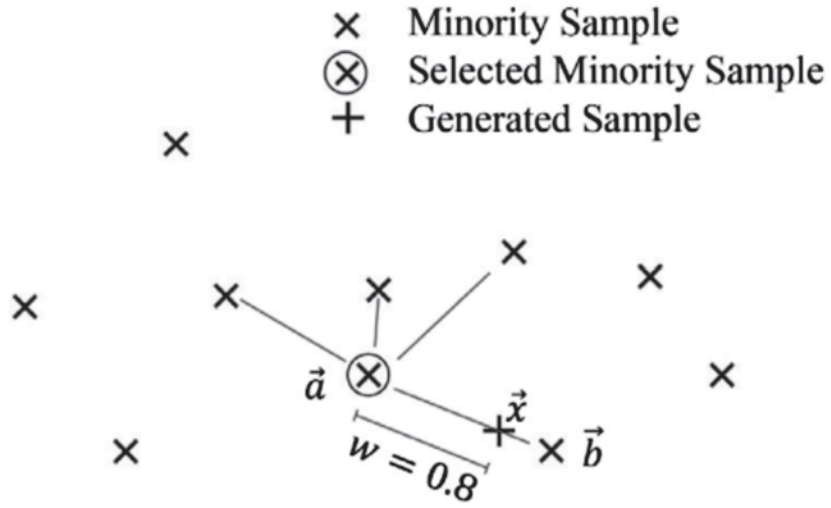

Figure 2. *SMOTE linearly interpolates a randomly selected minority sample and one of its k = 4 nearest neighbors*

5. MACHINE LEARNING ALGORITHMS

5.1. Random forest

The random forest algorithm is one of the supervised learning algorithms for classification and regression. A random forest is an ensemble method that includes multiple decision trees. This algorithm illustrates better outcomes once the number of trees is increasing in the forest and also prevents the model from overfitting. Each decision tree in the forest gives several outcomes. These outcomes are combined together so as to achieve more accurate and stable predictions [14 - 15].

5.2. K nearest neighbors

K nearest neighbors algorithm (KNN), is a non-parametric approach used to solve both types of problems and recovery processes. Despite the simplicity of this algorithm, it can do much better than a more systematic approach. KNN requests only the selection of *k*, the number of neighbors to be considered when implementing the classification. If the k value is small, the estimate of classification is prone to large statistical errors. On the contrary, if *k* is large in value, this allows the distant points to contribute to the classification, causing smooth out some of the details of the class distribution. Therefore, the value of *k* is considered to choose to minimize error classification on various independent number of data validation or by cross-validation procedures [16 - 17].

5.3. Decision tree

The decision tree (DT) algorithm is one of the most commonly used classifiers for classifying problems. This is mainly because this algorithm is capable of handling sophisticated problems by providing an understandable representation easier to interpret and also their adaptability to the inference task by generating logical rules of classification. A DT encompasses nodes for testing attributes, edges for branching by values of the selected attribute, and leaves labeling classes where for each leaf a unique class is attached. There are two major procedures in a DT that show one to build the tree, the other for the knowledge inference i.e. classification [18 - 19].

5.4. Logistic Regression

Logistic regression (LR) algorithm is one of the most prevailing approaches for binary classification. The relationship between predictors that might be continuous, binary, and categorical is indicated in the LR model. The dependent variable can be binary. According to several predictors, we anticipate whether something will occur or not. We identify the probability of belonging to each category for a given set of predictors [20].

5.5. Classification performance evaluation

In order to evaluate performances of the aforementioned ML algorithms on the resampled dataset, different criteria, named accuracy, recall, precision, and F1 score are chosen in this study [21]. These criteria are computed from a confusion matrix which is shown in **Table 1**. The confusion matrix is a good option to measure the performances of ML algorithms in the classification problems as it can observe the relations between the classifier outputs and the true ones. **Table 1** indicates four different combinations of predicted and actual values explained as follow:

a) True positive (TP): number of samples with absence of fraud predicted as absence of fraud.

b) False positive (FP): number of samples with presence of fraud predicted as absence of fraud.

c) True negative (TN): number of samples with presence of fraud predicted as presence of fraud.

d) False negative (FN): number of samples actually have absence of fraud predicted as presence of fraud.

Table 1. *Confusion matrix*

Actual Values	Predicted Values		
	Class	Negative (0)	Positive (1)
	Negative (0)	True Negative (TN)	False Positive (FP)
	Positive (1)	False Negative (FN)	True Positive (TP)

The classification measures are defined as below:

Classification accuracy is one of the most popularly employed measures for the performance of classification, and it is a ratio of the correctly classified samples over the total number of samples as shown in (1).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

Precision or positive predictive value illustrates the proportion of positive samples correctly classified to the total number of positive predicted samples as shown in(2).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall, sensitivity or true positive rate represents the ratio the proportion of positive samples correctly classified to the total of positive predicted samples and negative predicted samples as indicated in(3).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 measure is defined as the harmonic mean among precision and recall as illustrated in(4). F1 measure has a range from 0 to 1, which means its high values reveal high classification performance.

$$F_1 = \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

6. RESULTS

After using the SMOTE technique, the imbalanced dataset mentioned in section 2 achieved the balanced dataset as shown in **Figure 3.**

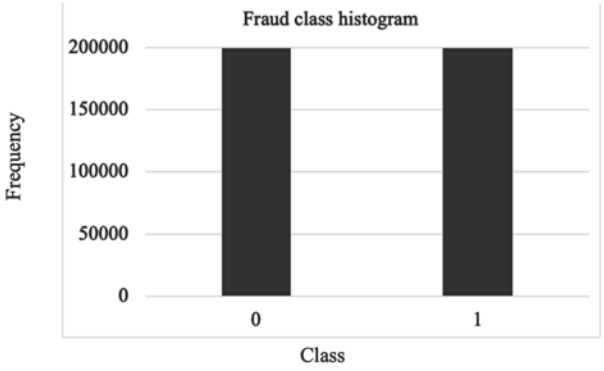

Figure 3. *Balanced dataset*

Additionally, the comparison of the accuracies of four machine learning approaches based on the imbalanced dataset and the balanced dataset obtained from the SMOTE technique is indicated in **Table 2**. The values of accuracies of the balanced dataset obtained from the SMOTE technique are higher than the ones of accuracies of the imbalanced dataset. As a result, the SMOTE technique demonstrated its effectiveness when generating the balanced dataset.

Table 2. *Comparison of accuracies of four machine learning approaches based on the imbalanced dataset and the balanced dataset obtained from the SMOTE technique*

ML approaches	Accuracies
RF with imbalanced dataset	96.12%
RF with balanced dataset based on SMOTE	99.95%
KNN with imbalanced dataset	95.72%
KNN with balanced dataset based on SMOTE	99.83%
DT with imbalanced dataset	94.89%
DT with balanced dataset based on SMOTE	99.76%
LR with imbalanced dataset	95.75%
LR with balanced dataset based on SMOTE	99.52%

Furthermore, the performance of four machine learning approaches based on the SMOTE technique is evaluated for the prediction of fraud using 30 features discussed in the dataset section. Through python 3.7 version, the confusion matrices of RF, KNN, DT, and LR algorithms based on the resampled dataset are accomplished to identify the TN, FN, FP, and TP. The confusion matrices of RF, KNN, DT, and LR algorithms with SMOTE are shown in **Table 3, Table 4, Table 5, Table 6** respectively.

Table 3. *Confusion matrix of RF with SMOTE*

True Values	Predicted Values		
	Class	0	1
	0	85284	12
	1	26	121

Table 4. *Confusion matrix of KNN with SMOTE*

True Values	Predicted Values		
	Class	0	1
	0	83878	1418
	1	44	103

Table 5. *Confusion matrix of Decision Tree with SMOTE*

True Values	Predicted Values		
	Class	0	1
	0	85123	173
	1	27	120

Table 6. *Confusion matrix of LR with SMOTE*

True Values	Predicted Values		
	Class	0	1
	0	83194	2102
	1	12	135

Classification performance measurements of four ML algorithms with SMOTE are revealed in **Table 7** and **Figure 3**.The accuracies of aforementioned algorithms with resampled dataset are almost similar (above 99%), except LR algorithms (approximately 97%). The precision of RF with SMOTE is about 92%, while this one of LR with SMOTE is only approximately 6%. Similarly, F1 score of RF with SMOTE is around 86%, while this one of LR with SMOTE is only 11%. Among four classification measures, the recall values do not indicate the big gap between ML algorithms relied on SMOTE. Moreover, the classification performance measurement of RF algorithm with SMOTE has generally better performances

compared with others, over 90% regard to accuracy and precision, and above 80% with recall and F1.

Table 7. *Classification measurements*

Measures	Random forest	KNN	Decision tree	Logistic regression
	SMOTE	SMOTE	SMOTE	SMOTE
Accuracy	99.95%	99.83%	99.76%	97.52%
Precision	92.24%	50.39%	40.95%	6.03%
Recall	80.95%	86.39%	81.63%	91.83%
F1	86.23%	63.65%	54.54%	11.32%

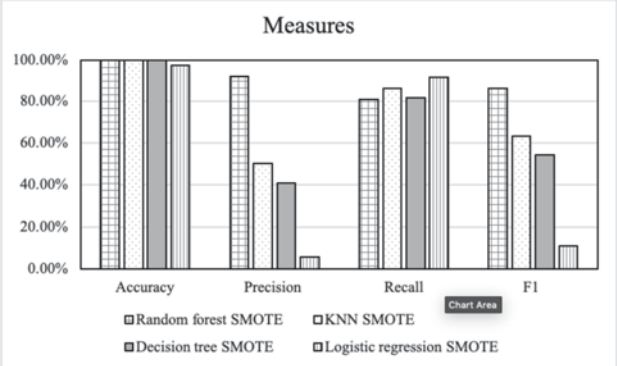

Figure 3. *Classification performance measurements*

7. CONCLUSION AND FUTURE WORK

In this study, performance evaluation of machine learning for prediction of credit fraud detection is implemented. We comprehensively analyzed the fraud dataset based on python version 3.7. The techniques such as SMOTE are applied for this dataset to achieve the balanced dataset. Four ML methods, then, are utilized to predict the detection of fraudulent activities via the balanced dataset. Four classification measurements, named accuracy, recall, precision, and F1 score which are calculated from confusion matrix are used to search for the suited performance for the problem of detecting fraud transactions. In the future, the dataset needs to collect precisely and extended through financial companies so as to accomplish the exhaustive model. Although the SMOTE technique is applied to generate the balanced dataset, the true positive samples are still less than the true negative ones. We continue to research other resampling techniques and ML algorithms to improve fraud detection system.

REFERENCES

[1] D. S. Sisodia, N. K. Reddy and S. Bhandari, "Performance evaluation of class balancing techniques for credit card fraud detection," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, 2017.

[2] B. Zhua, B. Baesens and S. K. Broucke, "An empirical comparison of techniques for the class imbalance problem in churn prediction," *Information Sciences*, vol. 408, pp. 84-99 , 2017.

[3] T. M. Padmaja, N. Dhulipalla, R. S. Bapi and P. Krishna, "Unbalanced Data Classification Using extreme outlier Elimination and Sampling Techniques for Fraud Detection," *15th International Conference on Advanced Computing and Communications (ADCOM)*, pp. 511-516, 2007.

[4] P. Kumari and S. P. Mishra, "Analysis of Credit Card Fraud Detection Using Fusion Classifiers," *Advances in Intelligent Systems and Computing*, vol. 711, pp. 111-122, 2018.

[5] R. Brause, T. Langsdorf and M. Hepp, "Neural data mining for credit card fraud detection," in *Proceedings 11th International Conference on Tools with Artificial Intelligence*, Chicago, IL, USA, 1999.

[6] A. Srivastava, A. Kundu, S. Sural and A. K. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model," *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING*, vol. 5, pp. 37 - 48, 2008.

[7] S. B. E. Raj and A. A. Portia, "Analysis on credit card fraud detection methods," in 2011 *International Conference on Computer,*

Communication and Electrical Technology (ICCCET), Tamilnadu, India, 2011.

[8] "Kaggle," [Online]. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud. [Accessed 2 9 2020].

[9] "Towards data science," [Online]. Available: https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02. [Accessed 5 9 2020].

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal Of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[11] K. Li, W. Zhang, Q. Lu and X. Fang, "An Improved SMOTE Imbalanced Data Classification Method Based on Support Degree," *in 2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, Beijing, China, 34-38.

[12] G. Douzas, F. Bacao and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, vol. 465, p. 1–20, 2018.

[13] T. K. Ho, "Random decision forests," in ICDAR '95: *Proceedings of the Third International Conference on Document Analysis and Recognition, IEEE Computer Society*, Washington, DC, USA, 1995.

[14] T. K. Ho, "Random decision forests," *in ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition, IEEE Computer Society*, Washington, DC, USA, 1995.

[15] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, p. 5–32, 2001.

[16] O. Beckonert, M. E. Bollard, T. M. Ebbels, H. C. Keun, H. Antti, E. Holmes, J. C. Lindon and J. K. Nicholson, "NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches," *Analytica Chimica Acta*, vol. 490, no. 1-2, p. 3–15, 2003.

[17] B. Alsbergav, R. Goodacrea, J. Rowlandb and D. Kella, "Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, K-nearest neighbour, neural and decision-tree methods," *Analytica Chimica Acta*, vol. 348, no. 1-3, pp. 389-407, 1997.

[18] J. R. Quinlan, "Induction of decision trees," *Machine Learning* , vol. 1, p. 81–106, 1986.

[19] J. R. Quinlan, C4.5 : programs for machine learning, San Mateo, Calif. : Morgan Kaufmann Publishers, 1993.

[20] Bishop and C. M, Pattern recognition and machine learning, New York: Springer, 2006.

[21] S.Kanmania, V. P.Thambiduraia, V.Sankaranarayanan and P.Thambiduraia, "Object-oriented software fault prediction using neural networks," *Information and Software Technology* , vol. 49, no. 5, pp. 483-492, 2007.

Phát hiện gian lận thẻ tín dụng dựa trên học máy

Trần Thành Công*

TÓM TẮT

Sự gia tăng các giao dịch gian lận thẻ tín dụng trực tuyến gây thiệt hại cho ngành tài chính. Hệ thống phát hiện gian lận, bao gồm hai nhãn gian lận và không gian lân cận được tạo ra. Tuy nhiên, dữ liệu này không cân bằng giữa hai lớp nhãn. Phương pháp lấy mẫu lại, SMOTE, được sử dụng để xử lý một tập dữ liệu không cân bằng thu được một tập dữ liệu cân bằng. Sau đó, các thuật toán học máy như rừng ngẫu nhiên, k hàng xóm gần nhất, cây quyết định và hồi quy logistic được áp dụng

cho tập dữ liệu cân bằng này để tạo ra các mô hình học máy để phát hiện gian lận thẻ. Hiệu suất của các môn hình học máy này được đánh giá bằng cách sử dụng độ chính xác, độ thu hồi, tính rõ ràng và điểm F1. Thuật toán rừng ngẫu nhiên dựa trên SMOTE xác định các gian lận theo cách tốt hơn các thuật toán khác.

***Từ khóa**: máy học, phân loại, dữ liệu không cân bằng, SMOTE, phát hiện gian lận*

Received: 17/12/2020
Revised: 28/12/2020
Accepted for publication: 31/12/2020