Detecting Financial Statement Fraud Using Random Forest with SMOTE

Huanzhuo Ye¹,*, Lin Xiang¹, and Yanping Gan²,b

¹Department of Information, Zhongnan University of Economics and Law, Wuhan, Hubei, China
²Wuhan University, Wuhan, Hubei, China

*Corresponding author e-mail: xianglinimi1204@foxmail.com,
yehuanzhuo@hotmail.com, yanping_gan@hotmail.com

Abstract. The study explores the comparison of various classification models in detecting fraudulent financial statements (FFS). Due to the high-class imbalance in this unique domain, the samples chosen in existing researches tend to be processed not so realistically. Therefore Random Forest is adopted to learn imbalanced data, in addition, sampling with SMOTE. Some more effective measure metrics of performance are also added. The experimental dataset includes 11726 publicly available Chinese financial disclosures from 2007 to 2017, of which 1314 financial statements were accused of fraud by CSRC. The result shows that the Random Forest outperforms other algorithms: Artificial Neural Network (ANN), Logistics Regression (LR), Support Vector Machines (SVM), CART, Decision Trees, Bayesian Networks, Bagging, Stacking and Adaboost.

1. Introduction

According to “The 2016 Report to the Nations on Occupational Fraud and Abuse” announced by ACFE [1], the total loss of fraud caused by the cases exceeded $6.3 billion, of which 23% of the cases exceeded $1 million. Financial statements Fraud accounted for less than 10% of fraud cases, but the median number of losses resulted in a loss of $ 975,000. Fraud perpetrators were inclined to expose behavioral warning signs (e.g. red flags [13]) when they were engaged in their crimes. Due to the higher transaction costs and the less efficient market caused by this fraud, financial statement fraud detection remains paramount to capture the potential criminal activity.

SAS No. 82 gives an explicit auditing definition that the auditor must provide “reasonable assurance about whether the financial statements are free of material misstatements, whether caused by error or fraud” [2]. Financial statement fraud usually means that corporations mislead investors and creditors by using falsified financial statements. It is generally known that financial statements can disclose information about the overall financial performance, like financial status, operating results and cash flow in the practice. Financial statement fraud detection can separate suspicious and authentic behavior in order to promote the development of targeted strategies.

In addition, fraudulent financial statements (FFS) have been a serious problem in China during the past decades, which significantly undermines the confidence of the investors. It is an inevitable problem because robust corporate governance is not perfectly developed. With the development of
data mining technology, researchers have also applied various techniques and models to detect FFS. In this stage, financial statement fraud detection has a great prospects in China. Many law enforcement and special investigative departments like China Securities Regulation Commission (CSRC) have also used Data Mining in daily work. Figure 1 shows a tendency of the numbers of published Chinese and English documents in authoritative journals on financial statement fraud during 1958-2016. It indicates that there is an increasing domestic research in recent years, even though it started later with fewer papers. The data are collected from Duxiu database which contains more than 430 million authoritative journals.

![Figure 1. A tendency of the numbers of published Chinese and English documents in authoritative journals on financial statement fraud during 1958-2016.](image)

Previous study has not effectively solved these two problems: missing values and severe class-imbalance. Firstly, samples tend to be removed with instances or features containing missing values. Listwise Deletion is a generally used method, simply removing the instance with incomplete information, which makes the conclusions biased or incomplete. Secondly, under certain circumstances, the prevalence of FFS can be tiny compared to legitimate (non-fraudulent). While under-sampling in the domain can cause biases in the dataset, the classifiers may be not robust in reality and inaccurately reveal real world scenarios. There is an urgent need to develop a model with realistic samples which depict the potential information of financial statement. The aim of the study is to use imputation during data preprocessing [25], given examples containing a high class-imbalance at a ratio of 1:7.9, and feature selection, to make the classification more robust in practice. It is proposed that Random Forest can be combined with SMOTE as a more appropriate alternative to the conventional ANN and logistics regression.

The paper proceeds as follows: Section 2 reviews the prior research in the domain. Section 3 describes the methodology of Random Forest in detecting financial statement fraud. In Section 4 a comparison of experiment results among Random Forest, ANN and logistic regression is presented, with respect to the performance metrics used in the study. Finally, a conclusion is made in Section 5.

2. Literature review

2.1. FFS detection

Traditional methods for detecting financial statement fraud are based on the prior knowledge of the auditors, which is strongly subjective. The accuracy rate and efficiency are rather low, despite it costs considerable time and labor. The steam of these methods focuses on statistical analysis and data mining classifications. Over the past decades, many data mining models have been proven to perform well in the domain. The two prevalent methods are logistic regression and different variations of ANN. Typically, logistic regression is used as a baseline model which the tested model is compared with. In this line of research, artificial neural networks is normally used, for example [10], [11], [14], [15], [17],
More recently, [16], [28] and [12] proposed genetic algorithm, CART and SVM, respectively. ID3, one kind of decision trees and Bayesian belief network were successfully used in [18]. Also, there is a trend of applying text mining methods [7] and [14].

Referring to using imbalanced samples, [16] applied the genetic algorithm on 390 US companies of 1:5.65. However, the model approximately correctly classified just 43% of the fraudulent companies. In [24], a logistic regression classification is developed on a dataset with an imbalance ratio of (4:21). The test divided the firms into “low risk, middle risk, high risk, very high risk”. The performance of model was unstable with the types, among which the best accuracy was 80.4%. Huang [17] compared two methods, SVM and LR on samples of 1: 4.8. It had the same unstable and low accuracy problems.

However almost all the prior works focused on comparisons among different models, which were based on an unrealistic prior fraud ratio of fraudulent to non-fraudulent firms (1:1). Prior fraud ratio (PFA) means the ratio of fraudulent to non-fraudulent firms of dataset. The model was not so stable and outperformed even though there are some study on imbalanced datasets. Furthermore the authors either just simply remove features or instances which contain missing values or have not mentioned the treatment at all. Specifically, small samples and fair prior fraud ratio (1:1) are two important reasons why previous study often had a high accuracy. Because most classifiers tend to achieve high accuracy for the majority class, but poor performance for the minority class. A variety of benchmarks have been used to measure classifiers’ performance, but the most common three standards used are accuracy, sensitivity, and specificity.

### 2.2. Imbalanced data learning approaches

Imbalanced data learning problem has been recognized in different application domains, such as credit card fraud detection, social sciences, taxes payment, and customer churn prediction, and customer retention, detection of oil spills from satellite images, medical diagnostic imaging, cancer detection and FFS and so on. Researchers have studied methods to deal with imbalanced problems, including five main categories [4]: sampling methods, cost-sensitive learning [31], ensemble learning methods, and feature selection methods and algorithms modification. First, sampling methods change the prior distribution of majority and minority class of the training set to achieve a more balanced number of cases in two classes. There are two common ways, over-sampling and under-sampling. SMOTE is one of the advanced sampling methods, along with Tomek Link [31], Neighbourhood Cleaning Rule NCL [21] and so on. Second, cost-sensitive learning [34] is mainly used in fraud detection [26], risk management or intrusion prevention. In such case, minimizing costs of the misclassification is the goal for the classifier developed. Third, ensemble learning methods include a series of boost classifiers, such as bagging, boosting and Random Forests. Fourth, feature selection methods can be a wrapper. The last one, algorithms modification means adjusting models in order to make them more appropriate in imbalanced dataset. Approach is often used with SVM [12] and Decision Tree.

Also, researchers have attempted to study to combine these methods. For example, [5] proposed a methodology using Random Forest with SMOTE when classifying an imbalanced multi-class dataset of UCI repository. But it didn’t work so well on high-dimensional data, only got 72% accuracy of Landsat dataset with 36 attributes. [30] Detected five attacks Intrusion Detection Systems (IDS) by sampling with SMOTE based on RFA. [33] Employed Random Forest with SMOTE in PAD Risk Factors Analysis, also using UCI repository. Referring to the dataset with 17 features and high imbalance, Adaboost is the best performer, however, not RFA. The above experiments are generally successful in their classification problems with an experimental dataset, but there is a lack of experiments to explore suitable sampling rate in SMOTE.

Few researchers have dealt with the imbalanced problem in the fraudulent financial statements detection, because the massive, multi-source and complex accounting data is a huge challenge. Financial data are required to be published publicly, and data from financial statements in different years are closely related to each other. Both reasons make it more difficult to distinguish the abnormal data from regulated data than general dataset. Compared with the dataset used in previous study, the sample in this study is more high-dimensional and comes from more than two sources. The
aforementioned study applies one-vs.-all to multi-class classification using Random Forest with SMOTE, which can lead to more biases during the process in sampling.

3. Methodology
Random Forest has been typically proven to be efficacious in learning imbalanced data [9]. So Random Forest algorithm is adopted to detect FFS. Because of the high class imbalance problem, in practice, it’s non-trivial when one only uses single simple classifier. Moreover, the evidence found here suggests that the Random Forest is inferior logistic regression for this problem. SMOTE is chosen to moderate biases caused by higher prior fraud probability. Also it is proposed to add performance metrics with precision, recall, F-Measure, Kappa Statistics and AUC when working with an imbalanced dataset, to be a supplement for accuracy.

3.1. Sample and Evaluation Measure
The original data contain data from 2011 Chinese companies’ publicly available financial disclosures from 2007 to 2017 in the A-share stock market. There are 11 industries divided by China Securities Regulatory Commission (CSRC) (see Table 1). In initial dataset, all financial statement fraud are reported from the fourth quarter of 2007 through the fourth quarter of 2017. For next step the dataset is reduced by eliminating: duplicates; non-annual financial statement fraud; financial companies. Financial firms are excluded due to the substantially different rules and regulations for financial companies from others. Non-annual financial statement frauds committed by firms are excluded because quarterly financial statements reveal information covering shorter time periods. As a result, 1314 financial disclosures have remained. 10412 non-fraud financial statements are matched randomly from other firms, which they do not appear in this set. The prior fraud ratio of final sample is approximately 1:7.9.

In particular, accuracy is not enough to work with an imbalanced dataset. The following performance metrics are used because they can give an insight into traditional classification accuracy (see Table 2). The average precision and recall over a 10-fold cross validation are followed.

| Industry Name | Code | Company Number |
|---------------|------|----------------|
| agriculture, forestry, husbandry and fishery | A | 51 |
| extractive industries | B | 84 |
| manufacturing | C | 2041 |
| production and supply of electricity, water and gas | D | 105 |
| construction | E | 101 |
| traffic transport and storage industry | F | 103 |
| information technology industry | G | 319 |
| wholesale and retail trades | H | 175 |
| real estate | I | 146 |
| social service industry | K | 137 |
| communication and cultural industry | L | 60 |
| cross-industry | M | 35 |

Table 1. All industries provided by CSMAR database
Figure 2. The pie chart of the industry distribution of FFS

Table 2. Metrics used and their meanings

| Metrics   | Meaning                                    |
|-----------|--------------------------------------------|
| Precision | classifiers exactness                      |
| Recall    | classifiers completeness                   |
| F-Measure | a weighted average of precision and recall |
| Kappa     | accuracy normalized by the imbalance of the classes in the data. |
| AUC       | area under the ROC curve                   |

3.2. Feature selection
Predicators are selected to be used as feature in the input vector combining with prior research relevant to FFS [3]. The predictors are leveraged to decide predictors in the study, which have been found to be comprehensive and significant, also easily available from CSMAR database. Table 4 shows the initial predictors including five aspects revealing debt paying ability, development capacity, risk level, dividend distribution, cash flow, profitability, management ability, internal controls. In total 25 financial ratios are compiled. To reduce dimensionality, chi-square is used to evaluate the worth of all predictors. 23 predictors are presented except for X6 and X11. The predictors were chosen to participate as the input vector. For further statement, features on internal controls are adopted because it is known that internal controls play an essential part in ensuring the authenticity and accuracy of accounting information. Fraud happens when there is deficiency in internal controls.
Table 3. Features chosen from different aspects in the field

| Code | Meaning                                      |
|------|----------------------------------------------|
| X1   | Current ratio                               |
| X2   | Quick ratio                                 |
| X3   | working capital                             |
| X4   | asset-liability ratio                       |
| X5   | Long-term Liabilities/ total assets         |
| X6   | Equity to debt ratio                        |
| X7   | Net profit growth rate                      |
| X8   | Total profit growth rate                    |
| X9   | Operating profit growth rate                |
| X10  | Revenue growth rate                         |
| X11  | Total revenue growth rate                   |
| X12  | financial leverage                          |
| X13  | Operating leverage                          |
| X14  | Degree of Total Leverage                    |
| X15  | Earnings retention                          |
| X16  | Net profit cash content                     |
| X17  | Net cash income from operating income       |
| X18  | operating cash index                        |
| X19  | EBIT                                        |
| X20  | Operating margin                            |
| X21  | Cash to profit ratio                        |
| X22  | Accounts receivable to income ratio         |
| X23  | Inventory to income ratio                   |
| X24  | Current assets to income ratio              |
| X25  | Is internal control deficiency?Yes:1, No: 0 |

3.3. Random Forest with SMOTE

There is a high imbalance problem in the training sample (minority: majority= 1: 7.9). Under the circumstances, traditional classifiers can be invalid, hard to recognize minority classes from the limited information. There are two types: under-sampling and over-sampling. An advanced method—SMOTE is employed due to the advantages of retaining some important information about majority class and making full use of existing information.

There are two common sampling methods to make a balanced dataset from an imbalanced dataset — under sampling and oversampling. Under sampling means removing some instances from the majority class, or pick partial instances from the majority class. Its main idea is keeping all instances in the minority class and randomly selecting an equal number of instances in the majority class to overcome the characteristics of algorithms. This method is used when quantity of data is sufficient. Under sampling can cause potential useful data loss so that the final result only has learnt part of the overall pattern. By contrast, over sampling duplicates instances of the minority class to increases the size of this class. The imbalanced sample may contain some repeated instances, which will lead to somewhat over-fitting.

Synthetic Minority Over-sampling Technique (SMOTE) is an over-sampling method proposed by Chawla Vet al. in 2002 [8]. It is an improved algorithm based on random sampling. Due to that the random sampling just simply copies the sample, it’s prone to over-fitting models, which means the information learned by the model is too specific (Specific) to be generalized (General). SMOTE’s main idea is to synthesize new minority class samples. The synthetic strategy is to select a sample B from each of its nearest neighbors, A, and then randomly select one between the lines of A and B as a new synthetic minority class sample. The algorithm is as follows:
(1) For each sample X in a minority classes, the distance from the Euclidean distance to all the samples in the training set is calculated and the k nearest neighbor is obtained.

(2) Setting a sampling ratio to determine the sampling rate N based on the sample imbalance ratio. For each minority class sample X, several samples are selected randomly from their k neighbors, assuming that the nearest neighbor is Xn.

(3) For each randomly selected neighbor Xn, constructing a new sample with the original sample according to the following formula.

In the given training sample of financial statement fraud, only 11.2% instances correspond to the positive class, such as, firms who committed FFS; the remaining majority instances belong to the negative class. RFA has become popular in classification field as an ensemble classifier method [6]. Therefore, it’s adopted to detect FFS for outstanding features of classifying simple and fast.

Random forest algorithm (RFA) is a versatile machine learning algorithm proposed by Leo Breiman [32] that can perform tasks of regression and classification. At the same time, it is a method for data dimensional reduction. It’s used to deal with outlier values, missing values and other data exploration. RFA is a bagging model with decision trees as base estimators. RAF has the following characteristics:

- Less parameter adjustment.
- Less probability to be over-fitting.
- Suitable to a large number of unknown features in a dataset.
- Possible to estimate features which are more important in classification.
- Good predication performance can be achieved in spite of a lot of noise in the dataset.
- Providing an effective way to balance the dataset errors dealing with an imbalance in classification.

RFA was constructed with multiple trees as opposed to just single one in CART or C4.5 model. To choose the final class based on input attributes, each tree gives a classification, which is called voting. In the end, RFA chooses the classification with the most votes. The forest takes the average of all outputs by each different tree for regression.

An explicit of RFA with SMOTE on how to work in detecting financial statement fraud is defined as follows.

Algorithm: RFA with SMOTE
Step 1: Sampling with SMOTE
  a. Cleaning the original dataset,
  b. Setting different sampling rate N based on the sample imbalance ratio of 1:7.9 to different new training samples.
  c. Setting Nearest Neighbors N (Default value is 5).
  d. Setting RandomSeed K (Default value is 1).
Step 2: Constructing Random Forest Classifier
  a. If there are N instances in initial training set S, using Bagging model to sample N instances from S as training set Si to grow the ith tree (Ti), then calling the CreatTree process to build decision tree.
  b. CreatTree can be divided into two steps:
    (1) If training set Si contains m dimensional features (Mall), then selecting Mtry randomly (Mtry Mall), then selecting the feature $\alpha$ with the best classification effect from Mtry, as the split feature of this node. Mtry is generally a constant. Here it takes Mtry as Int $[\log_2 (Mall) + 1]$.
    (2) According to feature $\alpha$, it splits this node into two branches, then calling CreatTree recursively for each branch, until the tree can accurately classify the training set, or all features have been used.
    (3) It is not pruned after the decision tree is built.
    (4) Repeat (1), (2) and (3) until the K decision tree was created.
Step 3: Using test set as the input of Step1 to train model, and finally the class labels (fraudulent vs non-fraudulent) of each instance are determined by voting.

RFA is a tree based model and it may causes bias and variance. Its construction tends to minimize the overall error rate, which focuses more on the accuracy of the majority class, as well as poor
accuracy of the minority class. As we can infer that a good model should maintain a balance between bias and variance errors. So SMOTE is employed for data preprocessing, which makes over-sampling and under-sampling combining. It can improve classifier performance and has advantages of these two traditional sampling methods [8].

4. Experiment Result and Analysis

4.1. Data pre-processing
In SMOTE, it is set that k=5, random seed=1, and N=500 based on the sample imbalance ratio of 1:7.9. As a result, a new training sample is generated that 7884 were fraudulent and 10412 were non-fraudulent where the prior fraud ratio is 1:1.32.

In addition, linear function conversion is used to normalize all numeric attribute values to [0,1]. To deal with missing values, missing numeric attributes are replaced by the mean, and missing nominal property is replaced with its mode. Replacing with missing value will be more suitable than Listwise Deletion. Figure 3 gives an insight of whole process of detecting FFS.

![Figure 3. Proposed Detecting FFS Framework](image)

4.2. Experimental Results
The original dataset includes 24 financial predictors for 2011 Chinese companies’ publicly available financial disclosures from 2007 to 2017 in the A-share stock market, of which 1314 were fraudulent and 10412 were non-fraudulent. Due to the high class imbalance in the dataset, performance metrics mentioned in Section 3 are chosen to take effect in the proposed combinatorial algorithm and other algorithms for comparison.

1) Result Summary
The results of the 10-fold cross-validation method for SVM, CART, C4.5, and Bayesian networks, ANN, logistic regression, bagging, stacking, Adaboost and RFA with feature selection on the same dataset were presented in Table 5. Precision means how many of the fraudulent financial statements classified were relevant. Recall is also called sensitivity which indicates how good a model is at detecting fraudulent financial statements. It is observed that RFA outperforms other classifiers with 87.8% precision and 87.5% recall (as indicated by italic numerals in Table 4), whereas CART came close behind with 76.5% precision and 77.2% recall. SVM has the lowest precision of 44.2% and recall of 66.5%. F-Measure is a metric interpreted as a weighted average of both precision and recall. It's also called F1 Score with the formula. The higher the F1 value on the interval [0, 1], the better the classification model works. It is observed that RFA has the highest F1 score 87.4%. Kappa Statistic is given by the formula, where is the observed accuracy and is the expected accuracy, which is based on
the marginal totals of the confusion matrix. The only model which achieves the equivalent Kappa (RFA) = 0.7418, which means the substantial consistency. The Area under the Curve (AUC) in Figure 4 illustrates the effectiveness of a binary classifier, (a) for RAF, (b) for LR and (c) for ANN in Figure 6. It shows the bad performance of ANN (AUC=0.688). A reason is maybe that there was over-fitting, and the performance of LR is worse (AUC=0.594). RAF has the highest AUC=0.9398, indicating its good performance.

![Figure 4. Receiver Operating Characteristic (ROC) curves of RFA, ANN and LR using the test dataset on same features.](image)

Considering the existing ensemble algorithms are generally successfully applied in FFS, three ensemble models are adopted to learn the same datasets, which are bagging, stacking, Adaboost. Bagging is a better performer producing precision of 81.5% and recall of 81.3% both higher than stacking and Adaboost. According to Kappa and AUC, bagging’s consistency performance is almost perfect and its AUC indicates better performance. Yet RFA also provides the best performance under the mentioned metrics. All models’ results are listed in Table 5 and the conclusion is consistent with that before. Performance of aforementioned classifiers listed with considerations of the metrics: RFA > Bagging > CART > C4.5 > ANN > Bayesian networks > Adaboost > LR > SVM > Stacking.

Table 4. Results of different models with reduced features (24 features selected by chi-statistic) and using 10-fold cross-validation

| Classifier         | Precision | Recall | F-Measure | AUC  | Kappa statistic |
|--------------------|-----------|--------|-----------|------|-----------------|
| LR                 | 64.2%     | 66.6%  | 54.1%     | 0.594| 0.0137          |
| ANN                | 73%       | 72.2%  | 67.2%     | 0.688| 0.2524          |
| SVM                | 44.2%     | 66.5%  | 53.1%     | 0.500| 0.0              |
| CART               | 76.5%     | 77.2%  | 76.3%     | 0.761| 0.4555          |
| C4.5               | 76.3%     | 76.8%  | 76.4%     | 0.739| 0.4648          |
| Bayesian- Networks | 72.4%     | 72.3%  | 72.3%     | 0.772| 0.3815          |
| Bagging            | 81.5%     | 81.3%  | 80.2%     | 0.860| 0.5451          |
| Stacking           | 44.2%     | 66.5%  | 53.1%     | 0.500| 0.0              |
| Adaboost           | 65.0%     | 67.7%  | 60.1%     | 0.666| 0.1035          |
| Random Forest      | 87.8%     | 87.5%  | 87.4%     | 0.940| 0.7418          |

2) RFA with SMOTE versus without SMOTE

Different prior fraud ratios are chosen in sampling techniques to comparatively evaluate RFA on the same training set. The percentages of SMOTE instances to create are 0, 100, 200, 300, 400 and
500. The prior fraud ratios after employing the SMOTE with different percentages is presented in Table 5. It can be concluded from the empirical results that RFA gets the best classification performance in FFS when the prior fraud ratio (PFA) is 1:1.32 with increasing percentage of 500. It is a paradox in the result of original dataset that RFA has the highest accuracy of 88.79 % while its AUC and Kappa are rather low. This may be accredited to over-fitting and unreliable measure of accuracy.

| Percentages | PFA  | Precision | Recall  | F-Measure | AUC  |
|------------|------|-----------|---------|-----------|------|
|            | 0    | 1:7.9     | 82.6%   | 88.8%     | 83.5%| 0.630|
| 100        | 1:3.96 | 86.4%     | 84.4%   | 80.3%     | 0.829|
| 200        | 1:2.64 | 85.9%     | 84.4%   | 82.4%     | 0.887|
| 300        | 1:1.98 | 86.9%     | 85.9%   | 85.1%     | 0.911|
| 400        | 1:1.58 | 87.3%     | 86.8%   | 86.4%     | 0.928|
| 500        | 1:1.32 | 87.8%     | 87.5%   | 87.4%     | 0.940|

5. Conclusions and Suggestions
This study has investigated the utilization of various Data Mining techniques in detecting financial statement fraud by using published financial disclosures. The methods employed were random forest algorithm, Artificial Neural Network, Logistics Regression, Support Vector Machine, CART, Decision Tree (C4.5), Bayesian Networks, Bagging, Stacking and Adaboost.

In summary, samples containing missing information have not been used effectively in detecting financial statement fraud. The experiment showed that SMOTE is a useful measure to sampling with the high class imbalance. Also, this study shows cost-sensitive classification like random forest algorithm will outperform the others with learning imbalanced dataset.

In terms of performance, the Random Forest Algorithm outperforms the well-known Artificial Neural Network (ANN) and Logistics Regression (LR) and other models. The precision manages to reach an average of 86.9%, while ANN and LR have average precisions of 73%, 64.5%, respectively. The type I error rate in RFA was the lowest among all models.

In order to extend the study, different datasets can be taken into account, imputation methods of missing values. In addition, the next study can enrich the predictors with non-structured data-qualitative information, for example, structure of the board of directors or more information about internal controls. Considering the evolutionary of FFS, there is an urgent need for us to combine with continuous audit in the process. Future research is needed to develop an effective and efficient real-time model, based on historical financial information, transferring identification after fraud to predicting ahead of possible fraudsters.

References
[1] ACFE, 2016, Report to the Nation on Occupational Fraud and Abuse, Association of Certified Fraud Examiners, Austin, TX
[2] AICPA, 1997, Statement on Auditing Standards (SAS) No. 82: Consideration of Fraud in a Financial Statement Audit, American Institute of Certified Public Accountants, New York, NY.
[3] Amara I, Amar A B, Jarboui A. Detection of Fraud in Financial Statements: French Companies as a Case Study [J]. International Journal of Academic Research in Accounting Finance & Management Sciences, 2013, 3 (3): 40-51.
[4] Bekkar M, (2009),“Développement d’un modèle de prédiction du churn clientèle en télécommunication”, Master 2 theis stat et économétrie, Unive Toulouse 1 Sciences Sociales.
[5] Bhagat R C, Patil S S. Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest [C] // Advance Computing Conference. IEEE, 2015.
[6] Bosch, A., Zisserman, X., Muñoz (2007), “Image Classification Using Random Forests and
Ferns”, IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil.

[7] Cecchini M, Aytug H, Koehler G J, et al. Making words work: Using financial text as a predictor of financial events [J]. Decision Support Systems, 2010, 50 (1): 164-175.

[8] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2011, 16 (1): 321-357.

[9] Chen C, Breiman L. Using Random Forest to Learn Imbalanced Data [J]. 2004.

[10] Deng Q. Application of support vector machine in the detection of fraudulent financial statements [C] // International Conference on Computer Science & Education. 2009: 1056 - 1059.

[11] Fanning KM, Cogger KO. Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management 1998; 7: 21–41.

[12] Farquad M A H, Bose I. Preprocessing unbalanced data using support vector machine [J]. Decision Support Systems, 2012, 53 (1): 226–233.

[13] Feroz, E., T. Kwon, V. Pastena, K. Park, 2000, “The Efficacy of Red-Flags in Predicting the SEC’s Targets: An Artificial Neural Networks Approach,” International Journal of Intelligent Systems in Accounting, Finance & Management 9 (3), pp. 145-157.

[14] Glancy FH, Yadav SB. A computational model for financial reporting fraud detection. Decision Support Systems 2011; 50: 595–601.

[15] Green BP, Choi JH. Assessing the risk of management fraud through neural network technology. Auditing: A Journal Of Practice & Theory 1997; 16: 14–28.

[16] Hoogs B, Kiehl T, Lacombe C, Senturk D. A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. Intelligent Systems in Accounting, Finance and Management 2007; 15: 41–56

[17] Huang SY. Fraud detection model by using support vector machine techniques. International Journal of Digital Content Technology and its Applications 2013; 7: 32-42.

[18] Kirkos E, Spathis C, Manolopoulos Y. Data Mining techniques for the detection of fraudulent financial statements [J]. Expert Systems with Applications, 2007, 32 (4): 995-1003.

[19] Kubat M., Holte R C., Matwin S., Kohavi R., Provost F., (1998), “Machine learning for the detection of oil spills in satellite radar images”, in Machine Learning, pp. 195–215.

[20] Larivière B and Van den Poel D, (2005), “Predicting customer retention and profitability by using random forests and regression forests techniques”, Expert Systems With Applications, 29

[21] Laurikkala J, (2001), “Improving Identification of Difficult Small Classes by Balancing Class Distribution”, AIME, LNAI 2101, pp. 63–66, Springer-Verlag Berlin Heidelberg.

[22] Lin, C.C., Chiu, A.A., Huang, S.Y., Yen, D.C.: Detecting the financial statement fraud: the analysis of the differences between data mining techniques and experts’ judgments. Knowl.-Based Syst. 89, 459–470 (2015)

[23] Lin, J., M. Hwang, J. Becker, 2003, “A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting,” Managerial Auditing Journal 18 (8), pp. 657-665.

[24] Lou Y I, Wang M L. Fraud Risk Factor Of The Fraud Triangle Assessing The Likelihood Of Fraudulent Financial Reporting [J]. Annals of Thoracic Surgery, 2011, 7 (2): 220-6.

[25] Moepya, S.O., Akhoury, S.S., Nelwamondo, F.V., Twala, B.: The role of imputation in detecting fraudulent financial reporting. Int. J. Innov. Comput. Inf. Control 12 (1), 333–356 (2016)

[26] Moepya, S.O., Akhoury, S.S., Nelwamondo,F.V.: Applying cost-sensitive classification for financial fraud detection under high class imbalance. In: 2014 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 183–192. IEEE (2014)

[27] Nicholas T. Using AdaBoost and Decision Stumps to Identify Spam E-mail[C]/ Natural Language Processing. 2003.

[28] Pradeep G, Ravi V, Nandan K, et al. Fraud Detection in Financial Statements Using
Evolutionary Computation Based Rule Miners [M] // Swarm, Evolutionary, and Memetic Computing. Springer International Publishing, 2014: 229-242.

[29] Ravisankar P, Ravi V, Rao G R, et al. Detection of financial statement fraud and feature selection using data mining techniques [J]. 2011, 50 (2): 491-500.

[30] Tesfahun A, Bhaskari D L. Intrusion Detection Using Random Forests Classifier with SMOTE and Feature Reduction [C] // International Conference on Cloud & Ubiquitous Computing & Emerging Technologies. IEEE, 2014: 127-132.

[31] Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. In The 2010 International joint conference on neural networks (IJCNN) (pp. 1–8, 18–23).

[32] Tomek Ivan, (1976) “An Experiment with the Edited Nearest-Neighbor Rule”, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 6, No. 6, pp. 448-452.

[33] Yao D, Yang J, Zhan X. An improved random forest algorithm for class-imbalanced data classification and its application in PAD risk factors analysis [J]. The Open Electrical & Electronic Engineering Journal, 2013, 7: 62-70.

[34] Zadrozny, B., Langford, J., Abe, N., (2003), “Cost-sensitive learning by Cost-Proporionate instance Weighting”, In Proceedings of the 3th International Conference on Data Mining.