FINANCIAL FRAUD DETECTION: A NEW ENSEMBLE LEARNING APPROACH FOR IMBALANCED DATA

Yiyang Bian  
*University of Science and Technology of China*, bianyiyang321@gmail.com

Min Cheng  
*City University of Hong Kong*, mc.cheng@my.cityu.edu.hk

Chen Yang  
*Shenzhen University*, yangc@szu.edu.cn

Yuan Yuan  
*City University of Hong Kong*, yyuan39-c@my.cityu.edu.hk

Qing Li  
*City University of Hong Kong*, qing.li@cityu.edu.hk

*See next page for additional authors*

Follow this and additional works at: [http://aisel.aisnet.org/pacis2016](http://aisel.aisnet.org/pacis2016)
FINANCIAL FRAUD DETECTION: A NEW ENSEMBLE LEARNING APPROACH FOR IMBALANCED DATA

Yiyang Bian, USTC-CityU Joint Advanced Research Center, University of Science and Technology of China, City University of Hong Kong, bianyiyang321@gmail.com
Min Chen, Department of Computer Science, City University of Hong Kong, Hong Kong, China, mc.cheng@my.cityu.edu.hk
Chen Yang, Shenzhen University, China, yangc@szu.edu.cn
Yuan Yuan, City University of Hong Kong, Hong Kong, China, yyuan39-c@my.cityu.edu.hk
Qing Li, Department of Computer Science, City University of Hong Kong, Hong Kong, China, qing.li@cityu.edu.hk
J. Leon Zhao, City University of Hong Kong, Hong Kong SAR, China, jlzhao@cityu.edu.hk
Liang Liang, University of Science and Technology of China, China, lliang@ustc.edu.cn

Abstract
With the rapid development of online and offline transactions, various financial fraud crimes happen every day. Financial fraud has seriously affected the health of economics and damaged the welfare of consumers, investors, as well as financial institutions. Prior studies apply several classification technologies, including decision trees, Bayesian networks, and support vector machines (SVM), to detect fraud detection. However, they ignore one important characteristic of fraud data, which is the number of valid records is largely smaller than the number of illegal fraud records. It implies the data is imbalanced. To resolve this issue, some researchers combine different sampling techniques to improve the detection accuracy of imbalanced fraud data. Among these techniques, ensemble learning is regarded as a perfect tool to handle the classification in imbalance data set. In this study, we propose a new ensemble method for financial fraud detection. This approach combines the bagging and boosting techniques together, in which the bagging technique can reduce the variance for the classification model through resampling the original data set, while boosting technique can reduce the bias of the model. In the future, we would conduct a series of experiments to evaluate the effectiveness of our approaches with the other state-of-the-art methods on real datasets.

Keywords: Financial fraud detection, Ensemble learning, Imbalanced data classification
1 INTRODUCTION

Nowadays, the online and offline trading increase rapidly. Especially thanks to the great advancements in IT technologies such as the networking technology, electronic payment technology and mobile computing services, there is an increasing emerge of electronic markets and online transactions. It has led to a growing number of financial fraud crimes (both online and offline) (Yeh et al. 2009). The fraud crimes have significant negative economic and social effects across the world. Financial fraud has seriously affected the market economic order and damaged the confidence and interest of consumers, investors and financial institutions. It also has resulted in a general loss of confidence in the integrity of businesses (Albrecht et al. 2008). Many investors, policy regulators and decision makers are trying to develop appropriate fraud protection strategies to decrease the impact of fraud and maintain the security of financial markets.

How to efficiently distinguish fraudulent financial data from authentic data is an everlasting challenge. Such detection methods are of interest to not only scholars but also standard setters, regulators, audit firms and investors (Cecchini et al. 2010). With the rapid development of business intelligence and big data analysis techniques, studies of financial fraud detection have gone into a new stage. From the data analysis perspective, the main aim of fraud detection is to use data mining algorithms to identify fraud or anomaly patterns in massive amounts of financial transactions records. To achieve this target, many researchers proposed innovative methods, algorithms and detection solutions. However, with the development of fraud detection methods, criminals of fraud have also been evolving their fraud approaches to avoid detection (Bolton et al. 2001). As Padmaja et al. (2007) mentioned, data mining techniques for fraud detection should be involved. Therefore, fraud detection methods need constant innovation (Bhattacharyya et al. 2011).

Fortunately, data mining techniques have already been proved to be useful in this domain such as credit card approval, bankruptcy prediction, and analysis of markets (Panigrahi et al. 2009). In prior financial fraud detection studies, the most commonly used classification methods are logistic regression and neural network approaches (Bhattacharyya et al. 2011; West et al. 2016a). Sohl et al. (1995) first applied neural network technology for financial statement fraud detection. In the same year, Persons (1995) employed logistic regression technique for fraud investigation and identification. Fanning et al. (1998) used a neural network based on published financial data to detect fraudulent financial statements. Bolton et al. (2002) provided us some statistical fraud detection methods based on statistical learning.

However, the performances of most existing methods are still less than ideal. Most of the existing methods for financial fraud detection are still unable to achieve adequate fraud detection capabilities (Abbasi et al. 2012). In consequence, financial fraud detection continues as an important challenge for business intelligence technologies.

The rest of this paper is organized as follows. In section 2 we present the related work in this field. Section 3 describes the framework of our proposed ensemble approach and selected features. Section 4 provides our evaluation methods and Section 5 concludes the paper.

2 RELATED WORKS

Financial fraud detection is a very hot research issue that has been studied by many researchers from both academic circles and industrial fields for decades. Recently, decision trees, Bayesian networks, and support vector machines (SVM) have been applied in studying this issue more frequently (Abbasi et al. 2012). For example, Kirkos et al. (2007) compared the financial fraud detection performance of decision trees, neural networks and Bayesian belief networks. Abbasi et al. (2012) and West et al. (2016a) summarized the existing classification methods for financial fraud detection comprehensively. West et al. (2016b) provided a review on key performance of classification metrics that used for
financial fraud detection. There are also some researchers tried to look this problem in a combinatorial perspective. Chan et al. (1999) tried to use scalable techniques to analyse massive amount of transaction data and they proposed a combining multiple learned fraud detectors under “cost model”. Bhattacharyya et al. (2011) discussed three techniques that employed in fraud detection study namely Logistic Regression, Support Vector Machines, and Random Forest. Another parts of researchers tried to seek a novel method. For instance, Padmaja et al. (2007) proposed a new method for fraud detection, which using extreme outlier elimination and k Reverse Nearest Neighbours.

With the development of BI technologies, the emergent detection methods may provide us some innovative way for predicting the occurrence of fraud (Anderson-Lehman et al. 2004). However, less of these detection methods talked about the difficulties to process imbalanced data. Actually, fraud detection is considered to be a similar classification problem but with a vast imbalance in fraudulent to transactions, and misclassifying problem (Duman et al. 2011). Financial fraud data being highly skewed or imbalanced is the norm (Phua et al. 2004).

Datasets usually exhibit imbalanced features in detection of financial fraud transactions. Usually, we treat majority class as the negative class and the minority as a positive class or interest class. How to deal with the imbalance data sets arises researchers’ interests due to the poor performance of traditional classifiers whose algorithms are designed only for balanced case (Jia, Zhang and He 2014). López et al. (2013) gave us several explanations about the raise of imbalanced data issues: (1) The training process of traditional classifiers is usually in favour of the majority class because it mainly uses Global Performance Measures such as accuracy to guide the training process; (2) The rules for minority class often discard in comparison with the rules that generated from majority class cause of the low coverage and mismatching rules; (3) The minority class may be treated as noise falsely however the real noise affect the identification of minority class contrarily.

Prior researchers did some works on the classification of imbalanced data. For example, Padmaja et al. (2007) combined different sampling techniques to improve the detection accuracy of imbalanced fraud data. Yen et al. (2009) mentioned in their research that many datasets in real applications such as fraud detection often have the imbalanced class distribution problem. They proposed a cluster-based under sampling approach for selecting the representative data as training data to improve the classification accuracy for minority class. To the best of our knowledge, the methods dealing with imbalanced data classification usually contain two levels, namely data level (mainly sampling) and algorithm level (including the cost-sensitive learning, ensemble algorithms and one-class classification), here we mainly introduce sampling and ensemble learning which are mainly two categories of methods to handle the classification in imbalance dataset:

Sampling Methods: The sampling methods usually contain two variants: over-sampling and under-sampling method. Specifically, Over-sampling consists of following variations: (1) random minority over-sampling with replacement, which is randomly over sample the minority class until it consists of as many samples as the majority class; (2) Synthetic Minority Over-Sampling Technique (SMOTE) (Chawla and Bowyer 2002). In this method, the minority class is “over-sampled” by creating “synthetic” examples rather than by over-sampling the original data, the advantage of SMOTE is that it will not result in over fitting when bringing additional information. Borderline-SMOTE (Han, Wang, and Mao 2005) and SVM-SMOTE (Tang, Zhang, and Chawla 2009) are two extensions of SMOTE, which only consider samples in some particular region. The under-sampling method includes (1) Random minority under-sampling with replacement (RSR); (2) Under Sampling with Cluster Centroids (USCC), which replacing majority class samples with cluster centres by K-Means algorithm. This method will perform outstanding if the majority classes are circle shaped; (3) Neighbourhood Cleaning Rule (NCL). This method suggested using the class of nearest neighbours to determine whether to remove the sample or not (Laurikkala 2001).

Ensemble Methods: Ensemble methods include bagging and boosting. Bagging (L. Breiman 1996), Random forest (L. Breiman 1999) and AdaBoost (Freund and Schapire 1996) are all reported ensemble learning methods to be successful in variance reduction. The main improvement of the
ensemble learning is the combination of multiple classifiers that could reduce the data variance. The idea of Bagging (Bootstrap Aggregating) is to split the majority class into multiple sub sets, and for each sub set, training a basic weak classifier together with the minority class, then integrate these classifiers into a strong classifier. Bagging is proposed to use random sampling technique to generate empirical distribution so as to approximate the real distribution of the data set. The AdaBoost (Adaptive Boosting) talks about the misclassified samples that generated by the former basic classifier will be augmented by assigning weights. The weighted data sets will be send to next basic classifier.

AdaBoost (Adaptive Boosting) focus on the misclassified samples that generated by the former basic classifier to be augmented by assigning weights. The weighted data sets will be send to next basic classifier to reduce the total bias error. The weighting strategy of AdaBoost is equivalent to resampling the data space (Sun et al.2007), which are applicable to most classification systems without changing their learning methods. Besides, it could eliminate the extra learning cost for exploring the optimal class distribution and representative samples (Ali et al. 2015). Moreover, compared with the method of eliminating samples from data set, it reduces the information loss, over fitting risk and bias error of a certain classification learning method (He and Garcia, 2009).

3  OUR APPROACH

In this section we will describe our proposed approach, which is designed for the imbalanced financial fraud data sets.

3.1  Financial fraud data and features

We collect and identify fraudulent financial data from the SEC Accounting and Auditing Enforcement Releases (AAERs) that were posted between 2009 and 2014. The data collection approach undertaken was consistent with prior studies (Abbasi et al. 2012; Cecchini et al. 2010). Firstly, we divide total data set into training and testing data by firm instance years. Secondly, we determine the ratio of non-financial fraud (nFF) data to financial fraud (FF) data in the training dataset and cluster the entire sample into some clusters. Thirdly, we randomly select the nFF class samples in each cluster and combine them and FF samples to obtain three training datasets. Finally, consistent with Yen et al. (2009) ’s work, we control the ratios of nFF class samples to FF class samples for $i_{th}$ cluster are 500/10, 300/50 and 200/40.

We select indicator ratios based on prior financial fraud detection studies and the generated feature set is shown in Table 1.

<table>
<thead>
<tr>
<th>Indicator ratios</th>
<th>Definition</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asset Quality Index</td>
<td>Ratio of non-current assets to total assets</td>
<td>(Beneish 1999);(Dikmen et al. 2010)</td>
</tr>
<tr>
<td>Depreciation Index</td>
<td>Ration of depreciation rate of period $t$ to last period $t-1$</td>
<td>(Beneish 1999);(Cecchini et al. 2010).</td>
</tr>
<tr>
<td>Gross Margin Index</td>
<td>Ratio of gross margin of period $t$ to last period $t-1$</td>
<td>(Beneish 1999);(Lin et al. 2003)</td>
</tr>
<tr>
<td>Cash Flow Earnings Difference</td>
<td>The impact of accruals on financial statements</td>
<td>(Beneish 1999);(Abbasi et al. 2012)</td>
</tr>
<tr>
<td>Sales Growth</td>
<td>Ration of Net sales in period $t$ to Net sales in period $t-1$</td>
<td>(Beneish 1999);(Dikmen et al. 2010); (Persons 1995); (Cecchini et al. 2010);</td>
</tr>
<tr>
<td>Asset Turnover</td>
<td>Ratio of sales to total assets</td>
<td>(Persons 1995);(Spathis 2002)</td>
</tr>
</tbody>
</table>
3.2 Bagging-based boosting framework

Follow the logic of ensemble approach, we combine the bagging and boosting technique together where the bagging can reduce the variance for the classification model through resampling the original financial fraud data set, and boosting can reduce the bias of the fraud classification framework. The main framework can be shown as follows:

As shown in the Figure 1, in the bagging process, we divide the training data set into two parts based on the financial fraud data (positive class) and non-fraud data (negative class) respectively, and then fit both classes into a boosting classifier. We use the asymmetric bagging to utilize the samples of financial fraud data as much as possible. The uneven split of the original data and bootstrap methods could ensure the diversity of the sub training data. For each sub-training data set, over-sampling (Random Oversampling/ SMOTE) and under-sampling (Random Under Sampling/ NCL) are used to balance the distribution of minority and majority classes. To reduce the variance and bias of the base learner, this model integrates Adaboost into each sub classifier. The final output is generated by the voting combination of each sub classifiers. The pseudo code of the framework is as follows:

Table 1. Selected indicators for feature set

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inventory Growth</td>
<td>Ratio of inventory growth in period ( t ) to period ( t-1 )</td>
<td>(Cecchini et al. 2010); (Dikmen et al. 2010)</td>
</tr>
<tr>
<td>Operating Performance Margin</td>
<td>Ratio of net profit to net sales</td>
<td>(Persons 1995); (Spathis 2002); (Cecchini et al. 2010)</td>
</tr>
<tr>
<td>Receivables Growth</td>
<td>Ratio of receivables in period ( t ) to receivables in period ( t-1 )</td>
<td>(Summers et al. 1998); (Cecchini et al. 2010)</td>
</tr>
<tr>
<td>Leverage</td>
<td>Ratio of total debt to total assets</td>
<td>(Beneish 1999); (Spathis 2002); (Cecchini et al. 2010)</td>
</tr>
</tbody>
</table>
Asymmetric Bagging-Boosting Algorithmic procedure

Input:

- D: Original Training Set; containing Positive Set P and Negative Set N;
- T: Number of Boosting iterations; B: Bootstrap Number;
- S: Sampling Method; I: Weak Learner

Step 1: bootstrap sampling

Draw |P| instances from Positive Set with replacement to form positive subset \( P_i \); with \( P_i \) and choose sampling method S to generate B training sets: \( X_1, X_2, \ldots, X_B \) from Negative Set N; for each \( X_i \): \( |X_i| = |P_i| + |N_i| \); \( |P_i| = |P| \); \( |N_i| \) is determined by sampling method S; (\( 1 \leq i \leq B \))

Step 2: boosting training

For each sub set \( X_i[\{(x_1, y_1), \ldots, (x_m, y_m)\}] \) where \( x_j \in X_i, y_j \in Y_i = \{-1, +1\} \)

Initialize \( W_i(j) = 1 / m \);

For \( t = 1, \ldots, T \)

Train weak learner using distribution \( W_i \)

Get weak hypothesis \( h_t : X_i \rightarrow \{-1,+1\} \) with error \( \varepsilon_t = \Pr_{j \sim W_i}[h_t(x_j) \neq y_j] \).

Choose \( \alpha_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right) \).

Update:

\[
W_{t+1}(j) = \frac{W_t(j) * e^{-\alpha_t}}{Z_t} \begin{cases} 
eq & \text{if } h_t(x_j) = y_j \\ = & \text{if } h_t(x_j) \neq y_j \end{cases}
\]

Where \( Z_t \) is a normalization factor (so as to let \( W_t \) be a distribution)

Output the final integrated hypothesis:

\[
H(x) = \text{sign} \left( \sum_{t=1}^{T} \sum_{i=1}^{B} \alpha_{t,i} h_{t,i}(x) \right)
\]

It is noted that the bagging technique can reduce the variance for the classification model through resampling the original data set, and boosting technique can reduce the bias of the model. Hence our proposed combined approach would benefit from both sides.
4 EVALUATION

Consistent with Hevner et al. (2004)’s research, we rigorously evaluated our proposed framework by conducting a series of experiments to assess the effectiveness of our approaches with the other methods (such as logistic regression, neural network, etc.) on synthetic datasets and real datasets. Traditionally, accuracy is the most commonly stander that used for the purposes. However, for imbalance data classification problem, accuracy is no longer a proper measurement criteria since the rare class has very little impact on accuracy as compared to the prevalent class. In this research, we choose the measurements of G-means and ROC to evaluate the framework.

\[
G\text{-mean} = \sqrt{\frac{TP\_rate \cdot TN\_rate}{TP + FP}}
\]

measures the balanced performance of a learning algorithm between these two classes. Where the precision and recall (true positive) are defined as:

\[
\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}.
\]

Precision is defined as Positive Predictive Value denoting the percentage of relevant objects that are identified for retrieval; Recall is defined as True Positive Rate denoting the percentage of retrieved objects that are relevant. And the true negative rate defined as:

\[
\text{TN\_rate} = \frac{TN}{TN + FP}
\]

Another measurement that we adopt is ROC (Receiver Operating Characteristic) instead of precision-recall curve since the ROC curve is robust to the class distribution of this data set. We compute the AUC (Area Under the ROC Curve) to evaluate the performance of our model.

5 DISCUSSION

Financial fraud detection approaches hardly can be confirmed without a full investigation (Bay et al. 2006). The effect of imbalanced classification strategies is largely ignored in previous financial detection research (Yen et al. 2009). This paper tries to propose an approach to handle the imbalanced financial data in fraud detection classification process. The expecting research contributins are as following: first, we provide a framework, which integrates BI methods into an ensemble approach with a predictable better accuracy and stability for financial datasets which contain more and disordered samples; second, we provide new insights to the understanding of the public financial fraud detection and BI literature; third, it is helpful to highlight the role of business intelligence analysis techniques in financial fraud detection domain.

Practically, investors, audit firms, and government regulators could also benefit from the advantage of asymmetric bagging and boosting to reduce the variance and bias of classification model of financial imbalanced data sets in the classification process. However since the minority class samples are used for each base learner, the limited number of minority class or the existence of noise data in minority class may deteriorate the performance of the final model when combining these classifiers.

In future research, the proposed framework will be improved by combining the existing methods with other competitive methods. The new framework should enable business intelligence techniques to perform at the best in the context of financial fraud detection.
References


