Researches of Detection of Fraudulent Financial Statements based on Data Mining

Dan Han

Henan Institute of Technology, Xinxiang 453000, China

Abstract

Financial statement fraud has been one of the biggest challenges in the modern business world. Financial accounting fraud detection (FAFD) has become an emerging topic of great importance for academic, research and industries. In this paper, the effectiveness of Data Mining (DM) classification techniques in detecting firms that issue fraudulent financial statements (FFS) and deals with the identification of factors associated to FFS are explored. Our study investigates the usefulness of Data Mining techniques including Decision Trees, Neural Networks and Bayesian Belief Networks in the identification of fraudulent financial statements. At last, we compare the three models in terms of their performances.

Keywords: Data mining, Fraudulent financial statements, Decision tree.

1. INTRODUCTION

Financial statement fraud is when corporations engage in certain practices designed to hide or maneuver the accounts of a corporation to help it continue to remain attractive to investors. A financial statement fraud may be actionable under both the False Claims Act and the Dodd Frank Act as well. You may have suffered a financial statement fraud or may have original information about a financial statement fraud, which means that you may be able to bring either a financial statement fraud lawsuit or a whistleblower lawsuit depending on the facts peculiar to your case.

The most common occurrence of financial statement fraud is when losses are underplayed or deliberately hidden by corporations. Financial statement fraud comprises deliberate misstatements or omissions of amounts or disclosures of financial statements to deceive financial statement users, particularly investors and creditors, outright falsification, alteration, or manipulation of material financial records, supporting documents, or business transactions, material intentional omissions or misrepresentations of events, transactions, accounts, or other significant information from which financial statements are prepared, deliberate misapplication of accounting principles, policies, and procedures used to measure, recognize, report, and disclose economic events and business transactions and also intentional omissions of disclosures or presentation of inadequate disclosures regarding accounting principles and policies and related financial amounts.

There are massive issues that emanate from financial statement fraud. Financial statement fraud undermines the reliability, quality, transparency, and integrity of the financial reporting process and jeopardizes the integrity and objectivity of the auditing profession, especially auditors and auditing firms. Financial statement fraud diminishes the confidence of the capital markets, as well as market participants, in the reliability of financial information and as a consequence makes the capital markets less efficient. In the bigger picture, it adversely affects the nation's economic growth and prosperity, results in huge litigation costs, destroys careers of individuals involved in financial statement fraud and causes bankruptcy or substantial economic losses by the company engaged in financial statement fraud. It causes devastation in the normal operations and performance of alleged companies and erodes public confidence and trust in the accounting and auditing profession. Ultimately financial statement fraud translates to massive stockholder losses and debts to creditors, not to mention emotional trauma to employees who lose their jobs and retirement funds.

Financial statement fraud may be committed by the senior and mid-level management of a corporation to fraudulently enhance the financial health of a business and enrich one's own net worth. Senior management may indulge in fraudulent cover-ups to exceed the earnings or revenue growth expectations of stock market, to comply with loan agreements, to increase the amount of financing available from asset-based loans and to meet...
a lender's criteria for granting/extending loan facilities. They may also fudge the statements to create a rosy picture for the shareholders.

How do we determine financial statement fraud? Financial statement fraud is designed to rob you of your hard earned earnings primarily to fulfill the ever growing corporate greed. Here are some of the clear red flags that you might want to consider. First and foremost, despite tight cash flows, the company will report profits which mean that gross profit levels will remain high despite pricing pressure. The statement will show that accounts receivable, accounts payable and stock levels are increasing even when sales are declining. Keep an eye out for payments as bonuses to senior management in a down economy. It also is indicative of financial statement fraud.

Management fraud can be defined as the deliberate fraud committed by management that causes damage to investors and creditors through material misleading financial statements. During the audit process, the auditors have to estimate the possibility of management fraud. In order to develop his or her expectations, the auditor employs analytical review techniques, which allow for the estimation of account balances without examining relevant individual transactions.

Fraser (Fraser et al., 1997) classifies analytical review techniques as non-quantitative, simple quantitative and advance quantitative. Advance quantitative techniques include sophisticated methods derived from statistics and artificial intelligence, like Neural Networks and regression analysis.

Data mining is known as gaining insights and identifying interesting patterns from the data stored in large databases in such a way that the patterns and insights are statistically reliable, previously unknown, and actionable. Data mining is also defined as “a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequently gaining knowledge from a large database”. The blending point between data mining and detecting accounting fraud is that, data mining as an advanced analytical tool may assist the auditors in decision making and detecting fraud. The data mining techniques have the potential to solve the contradiction between effect and efficiency of fraud detection. Data mining plays an important role in financial accounting fraud detection, as it is often applied to extract and discover the hidden patterns in very large collection of data. An auditor can never become certain about the legitimacy of and intention behind a fraudulent transaction. Concerning this reality, the most optimal and cost effective option is to find out enough evidences of fraud from the available data using specialized mathematical and data processing algorithms.

In our study, we carry out an in-depth examination of publicly available data from the financial statements of various firms in order to detect FFS by using Data Mining classification methods. The goal of this research is to identify the financial factors to be used by auditors in assessing the likelihood of FFS and introduce, apply, and evaluate the use of Data Mining methods in differentiating between fraud and non-fraud observations.

This work is organized as follows. Section 2 describes related work. In Section 3 we introduce common classification of Data Mining techniques for fraud detection. In Section 4 three methods are employed in our study are introduced. In Section 5 analysis of the experiment results is given. In Section 6 we conclude this paper.

2. RELATED WORK

The application of Data Mining techniques for financial classification is a rising research area. Many law enforcement and special investigative units, whose mission it is to identify fraudulent activities, have also used Data Mining successfully. Recent studies have attempted to build models that will predict the presence of management fraud. Bell (Bell et al., 2000) use a powerful generalized qualitative response model to predict management fraud based on a set of data developed by an international public accounting firm. Green (Green et al., 1997) developed a Neural Network fraud classification model. The model used five ratios and three accounts as input. The results showed that Neural Networks have significant capabilities when used as a fraud detection tool. Summers (Summers et al., 1998) constructed a cascaded logit model to investigate the relationship between insider trading and fraud. They found that, in the presence of fraud, insiders reduce their holdings of company stock through high levels of selling activity. Abbot (Abbot et al., 2000) employed statistical regression analysis to examine if the existence of an independent audit committee mitigates the likelihood of fraud. Spathis (Spathis 2002) constructed a model to detect falsified financial statements. He employed the statistical method of logistic
regression. The reported accuracy rate exceeded 84%. The results suggest that there is potential in detecting FFS through the analysis of published financial statements. Spathis and Doumpos (Spathis et al., 2002) used the UTADIS method to develop a falsified financial statement detection model. The method operates on the basis of a non-parametric regression-based framework. They also used the discriminant analysis and logit regression methods as benchmarks. Their results indicate that the UTADIS method performs better than the other statistical methods as regards the training and validation sample.

3. CLASSIFICATION OF DATA MINING TECHNIQUES FOR FRAUD DETECTION

In this section, a conceptual framework is proposed for the applications of data mining techniques to financial accounting fraud detection. The classification framework, which is shown in Figure 1, is based on the nature of data mining research and fraud detection research.

![Figure 1. The conceptual framework for data mining to FAFD](image)

Each of the six data mining application classes (classification, clustering, outlier detection, prediction, regression and visualization) is supported by a set of algorithms to extract the relevant relationships in the data.

Classification builds up (from the training set) and utilizes a model (on the target set) to predict the categorical labels of unknown objects to distinguish between objects of different classes (Zhang et al., 2004). Classification or prediction is the process of identifying a set of common features (patterns), and proposing models that describe and distinguish data classes or concepts. Common classification techniques include neural networks, the Naïve Bayes technique, decision trees and support vector machines.

Clustering is used to partition objects into previously unknown conceptually meaningful groups, with the objects in a cluster being similar to one another but very dissimilar to the objects in other clusters. Clustering is also known as data segmentation or partitioning and is regarded as a variant of unsupervised classification. Cluster analysis decomposes or partitions a data set (single or multivariate) into dissimilar groups so that the data points in one group are similar to each other and are as different as possible from the data points in other groups. It is suggested that data objects in each cluster should have high intra-cluster similarity within the same cluster but should have low inter-cluster similarity to those in other clusters (Zhang et al., 2004). The most common clustering techniques are the K-nearest neighbour, the Naïve Bayes technique and self-organizing maps.

Prediction estimates numeric and ordered future values based on the patterns of a data set. It is noted that, for prediction, the attribute, for which the value being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered). Neural networks and logistic model prediction are the most commonly used prediction techniques.

Outlier detection is employed to measure the "distance" between data objects to detect those objects that are grossly different from or inconsistent with the remaining data set. Data that appear to have different
characteristics than the rest of the population are called outliers (Agyemang et al., 2006). A commonly used technique in outlier detection is the discounting learning algorithm (Yamanishi et al., 2004).

Regression is a statistical methodology used to reveal the relationship between one or more independent variables and a dependent variable. The regression technique is typically undertaken using such mathematical methods as logistic regression and linear regression, and it is used in the detection of credit card, crop and automobile insurance, and corporate fraud.

Visualization refers to the easily understandable presentation of data and to methodology that converts complicated data characteristics into clear patterns to allow users to view the complex patterns or relationships uncovered in the data mining process. Visualization is best used to deliver complex patterns through the clear presentation of data or functions.

4. METHODS OF IDENTIFYING FRAUDULENT FINANCIAL STATEMENTS

Identifying fraudulent financial statements can be regarded as a typical classification problem. Data Mining proposes several classification methods derived from the fields of statistics and artificial intelligence. Three methods are employed in our study. These methods are Decision Trees, Neural Networks and Bayesian Belief Networks.

A Decision Tree (DT) is a tree structure, where each node represents a test on an attribute and each branch represents an outcome of the test. In this way, the predictive model attempts to divide observations into mutually exclusive subgroups and is used for data mining and machine learning tasks (Kirkos et al., 2007). These trees can be planted via machine-learning-based algorithms such as the ID3, CART and C4.5.

The successive division of the sample may produce a large tree. Some of the tree’s branches may reflect anomalies in the training set, like false values or outliers. For that reason, tree pruning is required. Tree pruning involves the removal of splitting nodes in a way that does not significantly affect the model’s accuracy rate (Koh et al., 2004).

Neural Networks (NN) is a mature technology with an established theory and recognized application areas. A NN consists of a number of neurons. Associated with each connection is a numerical value, called "weight". Each neuron receives signals from connected neurons and the combined input signal is calculated. The neurons are arranged into layers. A layered network consists of at least an input (first) and an output (last) layer. Between the input and output layer there may exist one or more hidden layers.

After the network architecture is defined, the network must be trained. In backpropagation networks, a pattern is applied to the input layer and a final output is calculated at the output layer. The output is compared with the desired result and the errors are propagated backwards in the NN by tuning the weights of the connections. This process iterates until an acceptable error rate is reached.

Bayesian classification is based on the statistical theorem of Bayes. Bayes theorem provides a calculation for the posterior probability. According to Bayes theorem, if \( H \) is a hypothesis—such as the object \( X \) belongs to the class \( C \)—then the probability that the hypothesis holds is \( P(H|X) = (P(X|H) \ast P(H))/P(X) \).

If an object \( X \) belongs to one of \( i \) alternative classes, in order to classify the object a Bayesian classifier calculates the probabilities \( P(C_i|X) \) for all the possible classes \( C_i \) and assigns the object to the class with the maximum probability \( P(C_i|X) \).

Bayesian Belief Networks (BBN) allow for the representation of dependencies among subsets of attributes. A BBN is a directed acyclic graph, where each node represents an attribute and each arrow represents a probabilistic dependence. If an arrow is drawn from node \( A \) to node \( B \), then \( A \) is parent of \( B \) and \( B \) is a descendent of \( A \). In a Belief Network, each variable is conditional independent of its nondescendents, given its parents (Hansen et al., 1996). The network structure can be defined in advance or can be inferred from the data. For classification purposes one of the nodes can be defined as the class node. The network can calculate the probability of each alternative class.

5. ANALYSIS OF THE EXPERIMENT RESULTS
In our experiment, three alternative models were built, each based on a different method. First, the Decision Tree model was constructed. The model was built with confidence level 0.05. We used the whole sample as a training set. Figure 2 shows the constructed Decision Tree.

![Decision Tree](image)

Figure 2. The decision tree

The model was tested against the training sample and managed to correctly classify 73 cases (general performance 96%). More specifically, the Decision Tree correctly classified all the non-fraud cases (100%) and 35 out of the 38 fraud cases (92%).

In the second experiment, we constructed the Neural Network model. We build a multi-layer perceptron feed-forward Network. After testing a number of alternative designs and performing preliminary trainings, we chose a topology with one hidden layer containing five hidden nodes.

The selected network was trained by using the whole sample and was tested against the training set. The network succeeded in correctly classifying all the cases, thus achieving a performance of 100%. Unfortunately, the software did not provide a transparent interface to the synaptic weights of the connections and thus an estimation of the importance of each input variable was not possible.

In the third experiment, we developed a Bayesian Belief Network (BBN). The implemented algorithm belongs to the category of conditional independence test-based algorithms and does not require node ordering. In order to train the Belief network, we used the whole sample as a training set. After the training, the network was tested against the training set. The network correctly classified 72 cases (performance 95%). In particular, it correctly classified 37 fraud cases (97%) and 35 non-fraud cases (92%).

The performances of the three models on the training sample are shown in Table 1. The results indicate that the NN model is quite efficient in discriminating between FFS and non-FFS firms, followed by the BBN and ID3 models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fraud (%)</th>
<th>Non-fraud (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID3</td>
<td>92.1</td>
<td>100.0</td>
<td>96.2</td>
</tr>
<tr>
<td>NN</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>BBN</td>
<td>97.4</td>
<td>92.1</td>
<td>94.7</td>
</tr>
</tbody>
</table>

Table 1 Performance against the training set
There are several approaches in model validation like dividing the sample into training and a separate hold out sample, 10-fold cross validation and hold one out validation. We chose to follow a stratified 10-fold cross validation approach. In 10-fold cross validation, the sample is divided in 10-folds. For each fold, the model is trained by using the remaining nine folds and tested by using the hold out fold. Finally, the average performance is calculated. The 10-fold cross validation performances of the three models are shown in Table 2.

Table 2 10-fold cross validation performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Fraud (%)</th>
<th>Non-fraud (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID3</td>
<td>75.0</td>
<td>72.5</td>
<td>73.6</td>
</tr>
<tr>
<td>NN</td>
<td>82.5</td>
<td>77.5</td>
<td>80.0</td>
</tr>
<tr>
<td>BBN</td>
<td>91.7</td>
<td>88.9</td>
<td>90.3</td>
</tr>
</tbody>
</table>

As expected the accuracy rates are lower for the validation set than the accuracy rates for the training set. However, the performance of each of the three models is considerably different. The Decision Tree model, which manages to correctly classify 96% of the training set, presents a considerable decrease of its classification accuracy when tested against the validation sample. The model correctly classifies 73.6% of the total sample, 75% of the fraud cases and 72.5% of the non-fraud cases. The Neural Network model, which enjoys an absolute performance of 100% on the training set, manages to correctly classify 80% of the total validation sample, 82.5% of the fraud cases and 77.5% of the non-fraud cases. Finally, the Bayesian Belief Network model which has the lower accuracy for the training set succeeds in correctly classifying 91.7% of the fraud cases, 88.9% of the non-fraud cases and 90.3% of the total validation set.

In a comparative assessment of the models’ performance we can conclude that the Bayesian Belief Network outperforms the other two models and achieves outstanding classification accuracy. Neural Networks achieve a satisfactorily high performance. Finally, the Decision Tree’s performance is considered rather low.

In assessing the performance of a model, another important consideration is the Type I and Type II error rates. A Type I error is committed when a fraud company is classified as non-fraud. A Type II error is committed when a non-fraud company is classified as fraud. Type I and Type II errors have different costs. Classifying a fraud company as non-fraud may lead to incorrect decisions, which may cause serious economic damage. The misclassification of a non-fraud company may cause additional investigations at the expense of the required time. Although any model aims to reduce both Type I and Type II error rates, a model is supposed to be preferable when it presents a Type I error rate which is lower than its Type II error rate. In our experiments, all the models present lower Type I error rates. Neural Network presents the greatest difference between Type I and Type II error rates.

6. CONCLUSIONS

Auditing practices nowadays have to cope with an increasing number of management fraud cases. Data Mining techniques, which claim they have advanced classification and prediction capabilities, could facilitate auditors in accomplishing the task of management fraud detection. The aim of this paper has been to investigate the usefulness and compare the performance of three Data Mining techniques in detecting fraudulent financial statements. The experiment results agree with prior research results indicating that published financial statement data contains falsification indicators.

REFERENCES