Fraud Detection in Financial Statements Using Classification Algorithm

Vasudha Sarda¹, Prasham Sakaria², Prof Dharmeshkumar Mistry³

¹,²,³Computer Science, D. J. Sanghvi College of Engineering, Mumbai, India

Abstract—There have been many high profile companies whose fraudulent financial statement is being telecasted widely. Data mining techniques have been used to detect such fraud in the financial statements. These extensive techniques have mostly considered only the quantitative part of the financial statement like the financial ratios but there has been very less usage of the qualitative information present for classifying the financial statement as fraudulent. There is very little research on the analysis of text such as auditor’s comments or notes present in published reports. Mining of the textual data can be used to detect fraud using text mining techniques. Support Vector Machine algorithm is used to analyze whether internal audit reporting structure and internal audit sourcing arrangement affect the ability of the internal audit function to prevent financial statement fraud or detect if the statements of a company are fraudulent in the financial statement reporting process.

Keywords—SVM-Support Vector Machine, Financial Statement, Fraud Detection, Text Mining.

I. INTRODUCTION

Data mining techniques have been used for detection of fraud in many fields. There have been insurance frauds, corporate frauds or credit card frauds which have been very popular in the recent years. On the other hand there has been less work done on the other types of fraud like mortgage or commodity fraud. With the increase in financial statement fraud, detecting them has been a matter of concern.

Detection and classification of the fraudulent data in financial statements has been done using data mining techniques like logistic regression, neural networks, CART and Naïve Bayes theorem.

The companies nowadays are generally affected by fraud and in particular are affected by financial statement fraud. The image of the company is tarnished by the news by the reveal of fraud in their financial statement. A live example being the Satyam scam which was in news for a long time in India. Such news not only affects the company under scrutiny but also the financial reporting process being used whose quality and integrity becomes vulnerable. The quality and effectiveness of the financial audits along with the reliability of the process has been undermined after the fraudulent activities like Enron Debacle.

II. METHODOLOGY

Data mining methods are used for not only preventing fraudulent financial reporting but also the detection of the fraud if it occurs. If decision tree algorithm is used for detection then the rules generated can be used for reporting fraud. Association rules produced at the end of the evaluation can be used by practitioners and researchers.

A. SVM-Support Vector Machine

Support vector machine which is a supervised learning model in machine learning, is used. It basically draws a hyperplane in an infinite dimension hyperplane. It is an associated learning algorithm which can be used for recognizing patterns and analyse data used for classification and regression. A set of training data is divided into one of the two categories thus building a classification model. The model then helps to classify the new data into one of the two categories in making it a non-probabilistic binary linear classifier. The model consists of dots in space along with a hyperplane that divides the space into two categories. A good separation between the two categories is achieved when the hyperplane has largest distance to the nearest training data point of any class.

III. LITERATURE REVIEW

Various data mining techniques have been used in the past by different researchers to detect fraud. Like Koh and Low had used decision tree, classification method to construct a model using data sample from 165 organizations. It included attributes such as assets to current liabilities, market value of equity to total assets, interest payments to earnings before interest and tax, net income to total assets, and retained earnings to total assets. Cecchini M. in 2005 examined quantitative variables along with text information for detection of fraud. The account ratios as well as the changes seen in the past few years were taken into account along with the qualitative information.

In 2007, Kirkos et al investigated the usefulness of three Data Mining classification methods namely Decision Trees, Neural Networks and Naïve Bayes by analyzing 27 financial ratios extracted from publicly available data of 76 Greek manufacturing firms for detecting fraudulent financial statements.
Further, Hoogs et al had created a genetic algorithm approach for detecting financial statement fraud by analyzing different comparative metrics, depending on specific financial metrics and ratios which would outline company’s performance.

Six data mining techniques namely Multilayer Feed Forward Neural Network (MLFF), Support Vector Machines (SVM), Genetic Programming (GP), Group Method of Data Handling (GMDH), Logistic Regression (LR), and Probabilistic Neural Network (PNN) were applied to Chinese companies by Ravishankar et al in 2011. It also included 35 variables and ratios taken from the publically available financial statements.

Financial Accounting Fraud Detection has received considerable attention from the investors, academic researchers, media, the financial community and regulators. Recently high profile financial frauds found at large companies like Enron, Lucent, WorldCom and Satyam, has increased the requirement of detecting and reporting financial fraud.

IV. EXISTING SYSTEM

An example of a classification problem is identifying fraud in a financial statement. The classification process is divided into two steps, where in the first step the model is trained using a sample data. The sample data is organized using tuples and attributes. The attributes belong to the class label and each attribute falls under each tuple which is known as supervised learning. In the second step the performance of the model is found out by classifying the object not present in the training and validation set.

The classification methods from the fields like statistical analysis and artificial intelligence are generally proposed for this problem. The data mining methods generally used for classification are Neural Network, Decision Tree and Naïve Bayes theorem.

The input used in the algorithm is composed of ratios derived from financial statements such as balance sheets and income statements to predict the accuracy of the three classification methods. The three models are compared in terms of their performances and accuracy. The result of the experiment shall find out the most effective algorithm out of the three along with that states the importance of variables in the detection process.

The fraudulent financial statements supposedly had more activation languages, group references and less usage of lexical diversity as compared to a non fraudulent one. The language used in the fraud financial reports seemed to appear credible by not talking much about the actual content.

The classification model which constitutes of both C4.5 and Naïve Bayes theorem has the highest accuracy. Results support the potential use of linguistic analyses by auditors to flag questionable financial disclosures and to assess fraud risk under Statement on Auditing Standards. The variables used as the input vector for the model was based on the previous research work done in the field of financial statement fraud detection.

The conclusion to the comparison between the three models states that the performance of the Bayesian theorem was highest as it had the maximum number of correctly classified instances as compared to the other two. Decision tree had the lowest performance while Neural Network had comparatively better performance.

V. EXPERIMENT

Text Mining techniques are applied on financial reports which are gathered from both a fraud and non fraud company of the same size. Companies with fraudulent history can be identified by analysing AAER’s issued by SEC. For information gathered from a fraud organization there will be information from a similar non fraud company.

It is followed by pre-processing step which involves extraction of text from financial statements and form a standard data on which text analysis can be applied. During pre-processing, the text is converted into lower case, punctuations are removed and spaces are removed. The occurrences of these words are very important and the information regarding this is used for training the classifier. The vector associated with each document is compared with typical vector associated with a given class. Similar vectors are used to find similar statements and used to differentiate between different ones. The vector spaces generated above will be used by next step for classifying organizations into fraud or non fraud. Support Vector Machine is used for classifying the data into two categories, fraud or non fraud after constructing a hyperplane. The evaluation of the accuracy of the technique is done using measures such as accuracy, precision and recall. We present a text mining approach for differentiating between fraud and non – fraud financial statements.

Semantic patterns are detected from the model which can help differentiate fraud statement to a non-fraud one. The effectiveness of integrated language model in detection of fraudulent financial fraud is verified from the empirical point of view.
The model developed in this paper can benefit market regulators and investors by successfully detecting financial statement fraud. An effective regulation mechanism can be created by looking at the textual context more like the patterns.

<table>
<thead>
<tr>
<th>Unqualified Report</th>
<th>Qualified Opinion</th>
<th>Adverse Opinion</th>
<th>Disclaimer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepared Using GAAP</td>
<td>Had we been able to obtain evidence</td>
<td>The financial statements do not present fairly</td>
<td>Adversely, we do not express an opinion</td>
</tr>
<tr>
<td>Comply with relevant statutory requirements and regulations</td>
<td>Note scope limitation</td>
<td>Because of the effects of the non-GAAP item</td>
<td>The scope was not sufficient</td>
</tr>
<tr>
<td>Financial statements referred to above present fairly</td>
<td>Except for the effects of not</td>
<td>Provide amounts if reasonably possible</td>
<td>We were unable to audit</td>
</tr>
</tbody>
</table>

The independent auditors report can be of four types which are considered as the training cases in support vector machine:

1. **Unqualified Report**: An opinion is said to be unqualified when the Auditor concludes that the Financial Statements give a true and fair view in accordance with the financial reporting framework used for the preparation and presentation of the Financial Statements. An Auditor gives a clean opinion or Unqualified Opinion when he or she does not have any significant reservation in respect of matters contained in the Financial Statements.

2. **Qualified Opinion**: Qualified report is issued when the auditor encountered one of two types of situations which do not comply with generally accepted accounting principles, however the rest of the financial statements are fairly presented.

3. **Disclaimer of Opinion**: It is commonly referred to simply as a **Disclaimer**, is issued when the auditor could not form and consequently refuses to present an opinion on the financial statements. This type of report is issued when the auditor tried to audit an entity but could not complete the work due to various reasons and does not issue an opinion.

4. **Adverse Opinion Report** is issued on the financial statements of a company when the financial statements are materially misstated and such misstatements have pervasive effect on the financial statements.

The occurrences of any of the above statements in the financial report are noted and a weight is given to the statements depending on their occurrence. The unqualified statement is given weight 1 as the statement is non fraud. The qualified and adverse opinion statements are given weights -1 as these statements indicate fraud in the reports while disclaimer is given a weight of 0. When the bag of words collects the occurrences of the respective statements and finds the total weight of a single financial report. If the total weight is positive it means that the statement is non fraud while if the weight is negative it is declared fraud using support vector machine technique.

VI. **FUTURE WORK AND CONCLUSION**

The accuracy could be increased by adding input vector with qualitative information, such as previous auditors' qualifications or the composition of the administrative board. An indicator can be calculated for a specific industry when it is studied in detail. These informative variables are being used for implementing association rule mining for prevention and three predictive mining techniques namely Decision Tree, Naïve Bayesian Classifier, Genetic programming for detection of financial statement fraud. SFS is employed to overcome information overload problems, and the SVM technique is then used to assess the likelihood of FFS. Particle swarm optimization technique is used to find the variables in Support Vector Machine method. CART or Decision Tree is generally used by the auditors to increase substantial testing during their audit procedures by adopting reliable and easy decision rules.

**REFERENCES**


