Int. J. of Advanced Networking and Applications
Volume: 01 Issue: 03 Pages: 159-163 (2009)

Financial Statement Fraud Detection by Data Mining

[*]G.Apparao, [**]Dr.Prof Arun Singh, *G.S.Rao, [*]B.Lalitha Bhavani, *K.Eswar, [***]D.Rajani

[*]GITAM University, [**]Magadh University, [***]GOVT. Polytechnic College for women

ABSTRACT

Financial losses due to financial statement frauds (FSF) are increasing day by day in the world. The industry recognizes the problem and is just now starting to act. Although prevention is the best way to reduce frauds, fraudsters are adaptive and will usually find ways to circumvent such measures. Detecting fraud is essential once prevention mechanism has failed. Several data mining algorithms have been developed that allow one to extract relevant knowledge from a large amount of data like fraudulent financial statements to detect FSF. It is an attempt to detect FSF; We present a generic framework to do our analysis.

Keywords: Financial fraud detection; fraudulent financial statements; data Mining; management fraud.

Date of Submission: October 20, 2009 Accepted : November 11, 2009

I. INTRODUCTION:

Financial statement frauds (FSF) have received considerable attention from the public, the financial community and regulatory bodies because of several high profile frauds reported at large corporations such as Enron, Lucent, and WorldCom and Satam computers over the last few years. Falsifying financial statements primarily consist of elements manipulating by overstating assets, profit, or understating liabilities. Detecting management fraud using normal audit procedures is a difficult task [12]. First, there is a shortage of knowledge concerning the characteristics of management fraud. Second, most auditors lack the experience necessary to detect it. Finally, financial managers and accountants are deliberately trying to deceive the auditors [16]. For such managers, who understand the limitations of an audit, standard auditing procedures may be insufficient. These limitations suggest the need for additional analytical procedures for the effective detection of false financial statements. Statistics and data mining methods have been applied successfully to detect activities such as money laundering, e-commerce credit card fraud, telecommunications fraud, insurance fraud, and computer intrusion etc. However, FSF is complicated and detecting them is difficult. People tend to question about how to do it and how effective they are. The main objective this paper is to provide a comprehensive review on financial fraud detection (FFD) process. Selected data-mining-based methods that have been used in FFD were examined.

II. RELATED WORK

A specific research community has spend a significant amount of effort in studying FFS from which a portfolio of data mining algorithms has been adopted for FFD. For instance, using a logit regression analysis, Beasley [3] found that no-fraud firms have boards with significantly higher percentages of outside members than fraud firms. Hansen et al. [19] used a powerful generalized qualitative response model to predict management fraud based on a set of data developed by an international public accounting firm. An experiment was conducted to examine the use of expert systems to enhance the performance of auditors [14]. Green and Choi [18] presented a neural network fraud classification model employing endogenous financial data. A classification model created from the learned behavior pattern is then applied to a test sample. Fanning and Cogger [16] also used an artificial neural network to predict management fraud. Using publicly available predictors of fraudulent financial statements, they found a model of eight variables with a high probability of detection. Beneish [7] investigated the incentives and the penalties related to earnings overstatements primarily in firms that are subject to accounting enforcement actions by the Securities and Exchange Commission. Abbott et al. [1] examined and measured the audit committee independence and activity in mitigating the likelihood of fraud.

Several researchers have attempted to synthesize the literature. For instance, Phua et al. [15] categorized, compared, and summarized from almost all published technical and review articles in automated fraud detection within the last 10 years. However, their research focuses on general detection such as terrorist detection, financial crime detection and intrusion and spam detection. In this study, we examine in-depth publicly available papers from the internet and journals about data mining and accounting for detecting FSF specially. We use 23 recent references (from years 1995 to 2008) about financial fraud detection methods and eight references about the relationship of auditor, governance and fraud as the basis for our research and analysis.

Volume: 01 Issue: 03 Pages: 159-163 (2009)

III. A CLASSIFICATION FRAMEWORK FOR FINANCIAL FRAUD DETECTION

Although many data mining algorithms have been adapted for fraud detection, their implementation still follows the traditional information flow of data mining - data collection, data integration, data preprocessing, data mining, and pattern evaluation. We expand the generic DM framework to consider specific characteristics of detection techniques for financial fraud (see Fig. 1).

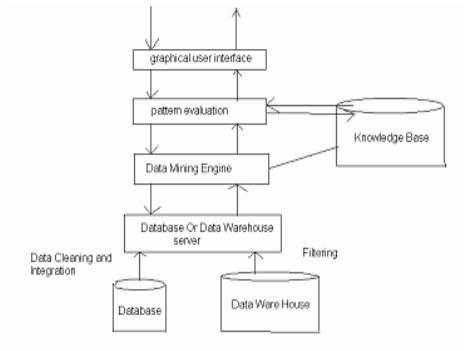

Figure 1: A Generic framework for DM -Based FFD

A. Data Distribution

The FFD algorithms can be first divided into two major categories, fraud & non-fraud company data and auditor data, based on the distribution of data. We summarize the literature according to data distribution in Table I.

TABLE I : SUMMARY BASED UPON DATA DISTRIBUTION

Data Distribution	Reference
Fraud company & non-fraud companies	[1, 3, 4, 5, 6, 7, 8, 13, 15, 16, 18, 19, 20, 21, 22, 24, 26, 27 29 30]
Auditor	[1, 8, 9, 14, 17, 23, 30]
Corporate governance	[3, 4]

As can be seen from Table I, earlier research has been predominately focused on dealing with fraud detection in combined fraud & non-fraud data. Abbott et al. [1] examined 41 firms which issued fraudulent reports and 88 firms which restated annual results without allegations of fraud in the period 1991-1999, together with matched pairs control groups of similar size, exchange listing, industry and auditor type. In Spathis's study [26], a sample of a total of 76 firms includes 38 with FFS and 38 non-FFS was examined. Ten financial variables are selected for examination as potential predictors of FFS.

The difficulties of applying FFD algorithms to other data can be attributed to two reasons: first, the auditor have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data for data mining, the fraud and non-fraud data, especially listed company, is easy to be obtained. Since today's financial fraud detecting techniques used to getting more difficult, using financial statement alone is insufficient to detect FFD. More attention and research should be focused on using fraud data with other information such as auditor and corporate governance.

B. Learning type

In Supervised machine learning the learning of the model is supervised in that it is told to which class each training sample belongs.. In other words, the goal of supervised learning is to build a concise model of the distribution of the class label in terms of the predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known but the value of the class label is unknown. Classification is learning by example. Unsupervised learning is another method of machine learning is grouping a set of physical or abstract objects in to classes of similar objects is called clustering. A cluster is collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters objects. Clustering is a learning by observation. We summarize the literature according to their learning type in Table II.

TABLE II : SUMMARY BASED UPON LEANING TYPE

FFD	Leaning Type	Reference
DM-based methods	Supervised	[1, 3, 5, 8, 13, 15, 16, 18, 19, 20, 21, 22, 24, 26, 27]
	Unsupervised	[4, 6, 7]

As can be seen from Table II, a majority of existing DM-based FFD algorithms used supervised learning method as a detection mechanism for mining fraud & non-fraud data. By using descriptive statistics, Beasley et al. [4] provides insight into financial statement fraud instances investigated during the late 1980s through the 1990s within three volatile industries— technology, health care, and financial services—and highlights important corporate governance differences between fraud companies and no-fraud benchmarks on an industry-by-industry basis. Beneish [7] uses same statistical method to investigate the incentives and the penalties related to earnings overstatements primarily in firms that are subject to accounting enforcement actions by the Securities and Exchange Commission. Unsupervised approaches have been used in outlier detection, spike detection, and other forms of scoring.

C. Data mining tasks/algorithms

The primitive data mining tasks which include classification, clustering, Association, Prediction and characterization. Currently, the FFD algorithms are mainly used on the tasks of classification. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the

purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups. We summarize the distribution of literature in Table

TABLE III : SUMMARY BASED UPON DETECTION APPROACH

Defection Approach	Detection Type	Frequency	%
Match analysis	Classification	26	100
	Clustering	0	0
Independent analysis	Clustering	0	0
	Time-series	0	0

As can be seen from Table III, currently only classification method has been used for mining fraud & non-fraud data. Compared with association rule mining, classification rule mining is more complicated to perform. Also, unlike association rules mining, which deals with existing data items, classification deals with attributes and its values. Moreover, instead of finding out the class label of attribute values, it also needs to step into fraud dataset and cluster the attributes further and make time-series mining or outlier detection for recognizing the new mode for detecting FFS with multi-firm-year feature.

D. Data mining technique

We can further divide FFD algorithms according to detection techniques used. Five techniques — regression, neural network, decision tree, Bayesian and SVM methodology — have been used to detect fraud data items for a data distribution centralized at one country. The idea behind regression is to establish a model using financial ratios from the firms to see which of the ratios were related to FFS. By including the data set of FFS and non-FFS we may find out which factors significantly influence the firms with FFS and then formulate the equation. The models will classify firms into FFS and non-FFS categories based upon financial statement ratios that have been documented as diagnostic in prior studies [26].

The SVM methodology revolves around the notion of a "margin" that separates two data classes. Maximizing the margin and thereby creating the largest possible distance between the separating hyperplanes can reduce the upper bound on the expected generalization error. However, most real-world problems involve non-separable data for which no hyperplane exists that successfully separates the positive from negative instances in the training set. The solution is then to map the data into a higher-dimensional space and define a separating hyperplane there. The distribution of the literature is given in Table IV.

As can be seen from Table IV, regression is the most popular method used, followed by artificial neural network. The regression models used include logit, stepwise-logistic, multicriteria decision aid method and exponential generalized beta two (EGB2) et al. Spathis et al. [27] used a sample of 76 firms, including 38 FFS and 38 non-FFS in Greece and ten financial variables, as potential predictors of FFS. They used univariate and multivariate statistical techniques such as logistic regression to develop a model to identify factors associated with FFS. A total often financial ratios are selected for examination as potential predictors of FFS. These variables appeared to be important in prior research and constitute ratios derived from published financial statements. The variables selected by the above techniques as possible indicators of FFS are: the inventories to sales ratio, the ratio of total debt to total assets, the working capital to total assets ratio, the net profit to total assets ratio, and financial distress (Z-score). Both models are accurate in classifying the total sample correctly with accuracy rates exceeding 84 per cent. The results of these models suggest that there is a good potential in detecting FFS through analysis of publicly available financial statements. In general the indicators selected are associated with FFS firms. Companies with high inventories with respect to sales, high debt to total assets, low net profit to total assets, low working capital to total assets and low Z scores are more likely to falsify financial statements according to the results of the stepwise logistic regression.

TABLE IV : SUMMARY BASED UPON DETECTION ALGORITHMS

Detection Algorithm	Reference
Regression	[1, 3, 5, 8, 19, 24, 26, 27 30]
Neural networks	[13, 15, 16, 18, 20, 22 27]
Statistical tests	[4, 6, 7 8]
Bayesian	[20]
Decision tree	[20]
Stacking variant methodology	[21]
Others non-DM based methods	[9]

The artificial neural network used includes not only generalized adaptive neural network architectures and the adaptive logic network but also fuzzy rule was integrated with a neural network [22]. Lin proved that the integrated fuzzy neural network outperformed most statistical models for neural networks reported in prior studies.

Only one study used three methods simultaneously, which include neural network, decision tree and Bayesian [20]. This study investigates the usefulness of these models in the identification of fraudulent financial statements. The input vector is composed of ratios derived from financial statements. The three models are compared in terms of their performances.

If one wants to obtain data mining results from data sources without class label, then the other method can be used like kmeans, genetic algorithms for clustering or time series analysis.

IV. SUGGESTIONS FOR FUTURE WORK

In this paper, we propose a generic FFD framework for understanding and classifying different

combinations of financial fraud detection techniques and data mining algorithms. The framework allows one to assess the different features of fraud detecting algorithms according to a variety of evaluation criteria. We examine 23 references to reveal current status of FFD'. The following directions were derived for future research.

First, feature selection is a very important stage in FFD. Currently there is no consensus on which data features are best for detection. Also, there is a need to combine financial data with other information such as auditor size, proportion and governance style for final analysis.

Second, most prior FFD algorithms were developed for use with fraud & non-fraud data simultaneously. However, with recent advances in fraud technologies, the more specific FFD methodology for fraud cases may have wider applications; especially we can combine multi-type data like financial ratio, auditor, governance and internal control for FFD.

Third, supervised learning techniques have been the dominated methods used for detecting FFS. However, those related algorithms do not pay full attention to new fraud features like over-cross several firm-years. Thus, further investigation, focusing on ensemble unsupervised and supervised learning mechanism will yield good results

Finally, selecting detecting algorithms for FFD has been a challenging yet unsolved issue. Future research can consider to propose an evaluation framework for common detection tasks, such as terrorist detection, financial crime detection and intrusion and spam detection.

REFERENCES

[1]. L.J. Abbott. S. Parker, and G F. Peters, G. F. "Audit committee characteristics and financial misstatement: A study of the efficacy of certain blue ribbon committee recommendation," Proceedings of the Auditing Section of the AAA Meeting, 2001.

[2]. A1CPA, Consideration of Fraud in a Financial Statement Audit, Statement on Auditing Standards No. 82, New York, NY: American Institute of Certified Public Accountants, 1997

[3]. M. S. Beasley, "An empirical analysis of the relation between the board of director composition and financial statement fraud," The Accounting Review, Vol. 71, No. 4. pp. 443-465, 1996.

[4]. M. S. Beasley. J. V. Carcello. D.R. Hermanson, and P. D. Lapides, "Fraudulent financial reporting: considerations of industrial traits and corporate governance mechanisms," Accounting Horizons, vol. 14, no. 4, pp. 441-454, 2000.

[5]. T. B. Bell, and J. V. Carcello, "A decision aid for assessing the likelihood of fraudulent financial reporting. Auditing: A Journal of Practice & Theory, Vol. 19, No. 1, pp. 169-184,2000,

[6]. M. D. Beneish, "Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance." Journal of Accounting and Public Policy, Vol. 16, No. 2, pp. 271-309, 1997.

[7]. M. D. Beneish, "Incentives and penalities related to earnings overstatements that violate GAAP," Accounting Review, vol. 4, no. 4, pp. 425-457. 1999

[8]. R. A Bernardi, "Fraud detection: the effect of client integrity and competence and auditor cognitive style," Auditing: A Journal of Practice & Theory. Supplement vol. 13, pp. 68-84, 1994.

[9]. J. R. Boatsman, C. Moeckel. and B. K. W. Pei, "The effects of decision consequences on auditors' reliance on decision dids in audit planning," Organizational Behavior and Human Decision Processes, vol. 71. pp. 211-247, 1997.

[10]. P. L. Brockett, R. Derrig, L, Golden, A. Levine, and M. Alpert, "Fraud classification using principal component analysis of RIDITs," The Journal of Risk and Insurance, Vol. 69, No. 3, pp. 341-371, 2002.

[11]. G. D. Coderre, Fraud Detection. Using Data Analysis Techniques to Detect Fraud, Vancover, Canada: Global Audit Publications, 1999.

[12]. F. Coglitore, and R. G. Berryman, "Analytical procedures: A defensive necessity," Auditing: A Journal of Practice & Theory, Vol. 7. No. 2, pp. 150-163, 1988.

[13]. E. H. Feroz, M. K. Taek, V. S. Pastena, and K. Park, "The Efficacy of red flags in predicting the see's targets: an artificial neural networks approach," International Journal of Intelligent Systems in Accounting, Finance & Management, Vol.9, pp. 145-157,2000.

[14]. M. M. Eining, DS. R. Jones, and J. K. Loebbecke, "Reliance on decision aids: an examination of auditors' assessment of management fraud." Auditing: A Journal of Practice and Theory, Vol. 16, pp. 1-19, 1997.

[15]. K. Fanning, K. Cogger, and R. Srivastava, "Detection of management fraud: a neural network approach", International Journal of Intelligent Systems in Accounting, Finance & Management, Vol. 4, No. 2, pp. 113-26, June 1995.

[16]. K. Fanning and K. Cogger, "Neural network detection of management fraud using published financial data," International Journal of Intelligent Systems in Accounting, Finance & Management, Vol. 7, No. 1, pp. 21-24, 1998.

[17]. B. P. Green, and T. G. Calderon, "Analytical procedures and auditors' capacity to detect management fraud," Accounting Enquiries, Vol. 5. No. 1, pp. 1-48, 1995.

[18]. B. P. Green, and J. H. Choi, "Assessing the risk of management fraud through neural network

technology," Auditing, Vol. 16, pp. 14-28, 1997.

[19]. J. V. Hansen, J. B. McDonald, and W. F. Messier, "A generalized qualitative-response model and the analysis of management fraud." Management Science, vol. 42, pp. 1022-1032, 1997.

[20]. E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," Expert Systems with Applications, pp. 32: 23, 2007

[21]. S. Kotsiantis, E. Koumanakos, D. Tzelepis, and V. Tampakas, "Forecasting fraudulent financial statements using data mining. International Journal of Computational Intelligence, Vol. 3, No. 2, pp. 104-110, 2006.

[22]. J. W. Lin, M. I. Hwang, and J. D. Becker, "A fuzzy neural network for assessing the risk of fraudulent financial reporting," Managerial Auditing Journal, Vol. 18, pp. 657-665,2003.

[23]. S. Owusu-Ansah, G. D. Moves, P. B. Oyelere, and D. Hay, "An empirical analysis of the likelihood of detecting fraud in New Zealand.," Managerial Auditing Journal, vol. 17, no. 4, pp. 192-204, 2002.

[24]. O. Persons, "Using financial statement data to identify factors associated with fraudulent financing reporting," Journal of Applied Business Research, vol. 11, pp. 38-46, 1995.

[25]. C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," Working paper, unpublished.

[26]. C. Spathis, "Detecting false financial statements using published, data: some evidence from Greece," Managerial Auditing Journal, vol. 17, no. 4, pp. 179-191,2002.

[27]. C. Spathis, M. Doumpos, and C. Zopounidis, "Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariates statistical techniques," European Accounting Review, Vol.11, pp. 506-535.2002

[28]. O. J. Welch, T. E. Reeves, and S. T Welch, Using a genetic algorithm-based classifier system for modeling auditor decision behavior in a fraud setting," Intelligent Systems in Accounting. Finance & Management , Vol.7, No. 3, pp" 173-186, 1998.

Authors Biography

Dr.G.Appa Rao., M.Tech., M.B.A., Ph.D., in computer science and Engineering form Andhra Universiy. Over 12 Years of teaching experience with GITAM University, handled courses for B.Tech, M.Tech. Research areas include Data Mining and AI. Published 8 papers in various National and International Conferences and Journals.

Dr.Prof Arun Singh,Ph.D., MAGADH UNIVERSITY, Dept of Computer Science and Engineering, over 15 years of teaching experience. Published 10 papers in various National and International Conferences and Journals.

Mr. G.Srinivasa Rao, M.Tech,(Ph.D)., Sr.Asst.Professor. He has Submitted Ph.D thesis in M.U., Over 7 Years of teaching experience with GITAM University, handled courses for B.Tech, M.Tech. Research areas include Computer Networks and Data Communications. Published 4 papers in various National and International Conferences and Journals.

Smt.D.Rajani.,M.Tech(AI&R),lecturer in Govt. Polytechnic for women, Bhimunipatnam, Over 9 years of teaching experience. Published 2 papers in various National and International Conferences and Journals.

K.Eswar M.Tech(IT) from GITAM University. Over 2 Years of teaching experience with AG college of engineering, Tadepalligudem, handled courses for B.Tech and M.C.A.

Ms.B.lalitha Bhavani M.Tech(IT) from GITAM University.