2010 International Conference on Intelligent Computation Technology and Automation

A Comprehensive Survey of Data Mining-based Accounting-Fraud Detection Research

Shiguo wang
Henan University of Science & Technology, Luoyang, HENAN, 471003, China
cicpa2003@163.com

Abstract—**This survey paper categorizes, compares, and summarizes the data set, algorithm and performance measurement in almost all published technical and review articles in automated accounting fraud detection. Most researches regard fraud companies and non-fraud companies as data subjects, Eigenvalue covers auditor data, company governance data, financial statement data, industries, trading data and other categories. Most data in earlier research were auditor data; Later research establish model by using sharing data and public statement data. Company governance data have been widely used. It is generally believed that ratio data is more effective than accounting data; Seldom research on time Series Data Mining were conducted. The retrieved literature used mining algorithms including statistical test, regression analysis, neural networks, decision tree, Bayesian network, and stack variables etc. Regression Analysis is widely used on hiding data. Generally the detecting effect and accuracy of NN are superior to regression model. General conclusion is that model detecting is better than auditor detecting rate without assisting. There is a need to introduce other algorithms of no-tag data mining. Owing to the small size of fraud samples, some literature reached conclusion based on training samples and may overestimated the effect of model.**

Data mining; Accounting-fraud detection; Data set; Algorithm; Classifier Evaluation

I. INTRODUCTION

Accounting fraud is intentional misstatements or omission of amount by deceiving users of financial statement, especially investors and creditors. Alrough the latest revision auditing standards is enlarging CPA's fraud-detection responsibility, effective detecting accounting fraud has always been a problem for accounting profession.

Data mining is about finding insights which are statistically reliable, unknown previously, and actionable from data (Elkan, 2001). The coupling point between data mining and detecting accounting fraud is that data mining as an advanced analytical program may assist decision making on auditor' detecting fraud and has the potential to solve the contradiction between effect and efficiency of fraud detection. It is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms. There are many researches that applied data mining algorithms on detecting accounting fraud. There are two kinds of critical suggestions concerning applying data mining technology on detecting fraud. One is lack of testable, open accessible data; The other is lack of mature methods and technologies (Clifton Phua, Vincent Lee, Kate Smith, Ross Gayler,2005). This paper aims to help certified public accountants selecting suitable data and data mining technologies on detecting fraud by reviewing literature on data structure, algorithms, conclusions, model effect evaluation etc., and provides reference for optimizing models.

II. LITERATURE REVIEW

Through analyzing fraud features of SEC accounting and AAERs, Loebbecke and Willingham proposed(1988) a basic model (L/W model) for detecting fraud which includes three categories of 46 risk factors:(1) conditions of organizations leading to fraud occurrence;(2)motivation of management authorities to fraud;(3)moral values of management authorities under the premise of recognizing the crime. By using univariate chi-square test, Loebbecke, Eining, Willingham(1989) tested detecting abilities of three categories in L/W model. It was concluded that 88% cases belonged to at least one of three categories of L/W model, Bell, Szykowny, Willingham (1991) continued to test L/W model by randomly selecting 500 samples of non-fraud cases. By applying Cascaded Logit model, there was more than 25% misreporting rate of three models, and among them, only 11% misreporting rate belonged to non-fraud cases by using optimized substitution model. Bell and Carcello (2000) furthered the research on detecting ability of fraud factors of L/W model by applying research samples of Loebbecke et al. (1989) and Bell et al. (1991), and they increased dummy variable's differentiate between earlier samples and recent samples and introduced ownership of samples into the model. The accuracy rate of prediction results of Logistic regression model was similar to Bell et al. (1991). Dummy variables test insignificant. The research compared discernment ability of model and certified public accountants and found out that model's differentiate on 77 cases of fraud samples was more accurate while there was no significant difference between model and professional auditors regarding to non-fraud samples.

For the first time, Calderon and Green(1994) established a fraud model by using sharing information. They used a major contradiction among revenue, accounts receivable and inventory of 111 companies of AAERs to test correlation of possible difference among prediction of financial analysts, profit of companies reports and fraud. The effect of model were as follows. When fraud really exists, the error risk of predicting non-existence of fraud increases from 9.76% to 15.38%.However, when fraud doesn't exist, the error risk of

978-0-7695-4077-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ICICTA.2010.831

predicting existence of fraud increases from 77.67% to 88%. Summers and Sweeney(1998)held that traditional red flag sign studied by Bell, Carcello(2000) was controllable and established model by creatively applying insider trading variables. And regression results of Cascaded Logit revealed that insider trading, the public information, can be used as signal of predicting fraud instead of traditional red flag sign. Persons (1995) provided to use open accessible report data to detect fraud. He selected fraud-related indexes from ten financial indexes representing seven aspects of companies. Final model prediction applied Jackknife method which calculated predicting scores of assessed companies, and classified assessed companies by comparing with benchmark points score making least cost of misclassification. Through calculating the cost of misclassification, the effect of model revealed that predicting results of model is superior to simple classification of all companies as non-fraud companies considering all the cost level different from type I error and type II error. Spathis(2002)established logistic regression model by using Greek data. There are similar researches like Beneish (1999) established a model detecting financial fraud including five financial indexes such as accounts receivable turnover index, gross margin index, asset quality index, sales growth index, total accruals to total assets index by using Probit method, and the predicting accuracy of the model reached 75%. Xuemin Huang(2007)concluded 22 financial indexes based on empirical analysis of horizontal(compared with the same industry average) and vertical(compared with the business of previous year)dimensions and intended to find out financial index which can significantly predict financial fraud. By regression analysis of using LPM model and Logit model separately, six financial indexes with high statistical significance were found, they are period fees-industry Index, gross margin index, depreciation-industry Index, non-recurring profit and loss index, recurring gains and losses - industry index, accounts receivable -trade index. Regarding to effect of model, Logit Model is superior to LPM model.

Beasley(1996) started to test creatively on standards of companies management supported by financial theory and policy-making organizations. And in the research, he tested the relationship between feature of management level in companies and incidence of fraud by using Logist regression. The research revealed that possibility of significant negative correlation between the proportion of outside directors and accounting fraud, existence and composition of auditing board didn’t significantly influence the occurrence of fraud; term increase of outsider directors in board, increase in the proportion holding , decrease in possibility of fraud Working in other companies; small size of board may decrease the possibility of fraud. However, the research didn’t provide predicting model. Regarding role of auditing board, there are different conclusions from researches of Abbott, Pareker, Peters(2002) , Beasley(2000) and Beasley(1996). Liguo Liu, Ying Du(2003) researched that there were a possibility of a positive correlation between financial fraud and the proportion of legal person shares and executive directors, inner control system, size of the board of supervisors and a negative correlation between the proportion of outstanding shares and them. Besides, there is a bigger chance for companies to fraud if the largest shareholder of the company is bureau of land and resources.

Green and Choi (1997) tested the validity of NN(Neural Networks) as screening tool. By literature retrieval, they established model by three different expectation methods of SPCNN, PSYDNN and ISYDNN regarding five ratios and three account variables as input variables(bad debts provision /net sales, bad debts provision/ accounts receivable, gross profit/net sales, accounts receivable/total assets, net sales, accounts receivable, bad debts provision). The classification ability of three NN were evaluated through learning, testing and blending samples. It was revealed that training effect of ratio data is superior to accounting data; NN enjoys a higher accuracy than other testing models, different from the initial expectation that more advanced ISYDNN may have a better effect than PSYDNN or SPCNN, increment-based expectation seems cover the model difference between fraud financial statement and non-fraud financial statement. Fanning And Cogger（1998）observed fraud model of managers based on ANN, and the model includes 8 detectable variables. Built a model of detecting accounting fraud based on FNN (Fuzzy NN)Lin, Hwang and Becker(2003)further proved that the effect of integrated FNN is much better than previous published statistics model of ANN, and the research compared the effect of integrated FNN with Baseline Logit, and through testing matching samples of fraud companies and non-fraud companies, he found that there was a highe accuracy for the two models on detecting non-fraud companies. And finally Logit model proposed by the research reached an accuracy rate of 97.5% on predicting non-fraud cases, which is much higher than other models with 86.7% accuracy rate.

Other technical researches include: Hansen et al.(1996)established detecting fraud model of managers by applying Probit and Logit based on data developed by international accounting firm; Beneish(1997)provided probability analysis to test fraud and profit control index proposed by the research model was obtained by combination of the value of the financial variables and eventually becoming possibility of profit control. Eining et al.(1997)examined the role of Expert Systems in increasing the detecting ability of auditors and statement users. By using expert system, they could have better detecting abilities to accounting fraud risk under different context and level and enable auditors give much reliable auditing suggestions through rational auditing procedure.

Relevant scholars in China mined detecting signal of domestic accounting fraud by using domestic data and tried to introduce new algorithm. Guoxin Chen, Zhanjia Lv, Feng He(2007)established fraud detecting model. Selecting 126 fraud companies and 126 normal companies from listed companies of Shenzhen and Shanghai Stock Exchange between 1994 and 2005 as samples, and selecting 29 indexes from categories such as finance, ownership structure, inner control and other four categories. Regression results showed that listed companies with low profit ability, high proportion of management ownership, few independent directors and lack of standard unqualified opinion has bigger chance to commit fraud. The model based on Logistic reached 95.1% of detecting accuracy with significant expectation effect. Based on the samples of 35 public published fraud companies and random 500 non-fraud companies from 2000 to 2004, Haisong Ren(2006)established detecting model of fraud from four aspects of financial indexes, company governance, financial risk and pressure and related trading by using clustering analysis and Logistic analysis. After cluster filtering significant variables, prediction model was established with methods of Standardization, non-Standardization Bayes and Logistic. The testing results revealed that the detecting results of four categories are basically similar, but detecting rate was very different. According to the detecting results of Bayes(Fisher),the highest rate of detecting fraud is financial risk and pressure, and the second is company governance, and the last is financial index. The detecting effect was not ideal with correlating trading. Regarding 76 Greek fraud and non-fraud companies as samples, Kirkos,Spathis and Manolopoulos(2007)established accounting fraud testing model by considering financial ratio as input variable and compared the effect of each detecting ability by using ANN, decision tree and Bayesian Network. And it showed that the effect of Bayesian Network was the best, and there was 90.3% of accuracy rate considering samples of 10-layer cross-validation; the accuracy rates of NN and decision tree were 80% and 73.6% respectively. The type I error rate of three models were all very low. Bayesian trust network revealed that there was a dependent relationship between debt conversion rate, ROA, sales to total assets ratio, working capital to total assets ratio, value Z and fraud.

III. SIMPLE CONCLUSION AND REVIEW

A. *Red Flag Sign of Accounting Fraud*

Most current researches regard fraud companies punished openly by securities regulatory authorities and non-fraud companies controlling industries, size and time as data subjects, Eigenvalue almost covers auditor data, company governance data, financial statement data, industries , trading data and other categories. And all these characteristic variables are concluded based on normative analysis, case analysis, empirical literature retrieval and reflects the basic characteristics of "triangle", and have strong correlation of accounting fraud and theory foundation. Earlier research regarding assisting decision-making of auditors as starting point, and most of data were auditor data, and there are several difficulties to apply in testing algorithm of accounting fraud: First, auditors are unwilling to publish personal concerned data; Second, although they want to provide data for data mining, it is difficult to collect the data; Third, data is very subjective. Later research tried to establish model by using sharing data, and public statement data, company governance data have been widely used. The research aim enlarged into not only assist auditors to make decisions, but also apply for regulation department and general investors. Current research focuses on reflecting comprehensive data of fraud triangle to face much more complicated fraud technologies.

The data mining results of existing literatures provided empirical evidence for "red flags" concluded from criteria and other theories researches .Some literatures even sort the sequence of fraud sings based on detecting ability. It is generally believed that ratio data is more effective than accounting data; there are several researched reached inconsistent conclusions on detecting ability test of auditing committee index due to the malfunction of earlier auditing committee; there was inconsistent conclusion between correlation trading and expectation, and research found that there are plenty of correlation trading between fraud companies and non-fraud companies so that the difference was covered.

One challenge on data for Chinese scholars is based on the popular of accounting fraud and low detection rate, current research using unpunished companies matched with known fraud companies, while unpunished companies doesn't mean without fraud. Therefore if there are undiscovered fraud companies matched with fraud companies, the reliability of data and the conclusion of data mining might be questioned.(Dianmin Yue,2007) Besides, the selections of matching companies only control industries and size variables, however, generally it is impossible to remove the difference influence between companies although applying matching comparative testing. Relatively, if using vertical comparison of the same companies during different periods can consider much more influential factors of companies itself, and may have better prediction effects. However, seldom research on time Series Data Mining was conducted.

B. *Algorithm*

The retrieved literature used mining algorithms including statistical test, regression analysis, NN, decision tree, Bayesian network, and stack variables etc. Regression Analysis is widely used on hiding data. Regression of fraud detecting has great explanation ability, and regression model used by literature are Logit, Step-wise Logistic, UTADIS and EGB2 etc, another is NN. Although it cannot be compared simply, generally the detecting effect and accuracy of NN are superior to regression model. The advantages of NN are that there are no strict requests for data and it has a strong generalization and adjustment. Although NN after correct allocation and training may make continually great classification and conclusion, comparing with regression model, due to special inner structure, it is impossible for researchers to track the formation process of output conclusion. And there has not clear explanation on connecting weight and cannot determine its accuracy and statistical reliability, and therefore lack of explanation.

Current algorithms are suitable for current literature with learning type of tags, and there is a need to introduce other algorithms if there has not tag data mining.

C. Classifier Evaluation

Generally based on classifier, correct signal proposition providing to auditors to allocate limited resources is used to evaluate the effect, and most of literature applies error scores standards to evaluate the effect of classifier. The general error rate of most of later literature on fraud detecting model is no more than 30% and reached a rather satisfied classifier accuracy. Related literature also compared the difference between model detecting rate and auditor detecting rate without assisting. General conclusion is that model detecting is better than auditor detecting rate without assisting. Bell&Carcello(2000) specially tested the detecting rate of model on the existence of fraud under any risk level is higher than auditor detecting rate without assisting, and it provided empirical evidence of fraud to improve auditors' estimate for priori probability.

Under the premise of any type I , type II error cost level, Persons(1995)calculated the lower error classification cost to evaluate effect of model and it was a special standards to evaluating model. There is a difference on evaluation methods of literature on error classification rate. Some literature differentiate training samples and testing samples, while owing to the small size of fraud samples, some literature reached conclusion based on training samples and may overestimated the effect of model. Under the small size of samples, Persons (1995) applied almost natural Jackknife which is a great method both maintaining training samples and improve credibility of testing.

REFERENCES

[1] Zhayi RuiFinancial Statements Fraud: Prevention and discovery[M].. Beijing: China Renmin University Press, 2005

[2] Elkan, C. (2001). Magical Thinking in Data Mining: Lessons from COIL Challenge 2000. *Proc. of SIGKDD01*, 426-431.

[3] Lavrac, N., Motoda, H., Fawcett, T., Holte, R., Langley, P. &Adriaans,. Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving.*Machine Learning* ,2004, (1-2)P. 57: 13-34.

[4] Ashutosh Deshimukh, Khondkar E.Karim, Philp H.Siegel. Analysis of Efficiency and Effectiveness of Auditing to Detect Management Fruad. *International Journal of Auditing,1998,2(2):127-138*

[5] Brian Patrick Green,Jae Hwa Choi. Assessing the Risk of Management Farud Through Neural Network Technology.*Auditing:A Jounrnal of Practice&Theory,1997,16(l):14-28*

[6] Clifton Phua, Vincent Lee, Kate Smith& Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *Working paper,2005*

[7] Loebbecke, J. K., M. and Willingham. 1988. .Review of SEC Accounting and Auditing Enforcement Releases. *Working Paper, University of Utah.*

[8] Loebbecke, J. K., M. Eining, and J. Willingham. 1989. Auditors' experience with material irregularities: Frequency, nature and detectability. *Auditing: A Journal of Practice & Theory* (Fall): 1-28.

[9] Bell, T. B., S. Szykowny, and J. J. Willingham. 1991. Assessing the likelihood of fraudulent financial reporting: A cascaded logit approach. *Working paper, KPMG LLP.*

[10] Bell T.B.,Carcello J.V. A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting[J]. *Auditing: A Journal of Practice &Theory, 2000,19(1):169-184.*

[11] Calderon, T. G., and B. P. Green. 1994. Analysts' forecast as an exogenous risk indicator in analytical auditing. *Advances in Accounting 12: 281-300.*

[12] Summers S.L., Sweeney J.T., Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis[J].*The Accounting Review,1998,73(1):131-146.*

[13] Persons O.S. Using Financial Statement Data to Identify Factors Associated with Fraudulent Financial Reporting[J].*Journal of Applied Business Research,1995,11(3):38-46.*

[14] Charalambos T. Spathis. Detecting False Financial Statements using Published Data: Some Evidence from Greece[J].*Managerial Auditing Journal,2002,17(4):179-191.*

[15] Beneish M D. Incentives and Penalties Related to Earnings Overstatements That Violate GAAP[J].*The Accounting Review,1999,74(4):425-457.*

[16] Xuemin Huang, Research on Public Company Accounting Fraud and regulation-from perspective of protecting investors[D]. Xiamen: Xiamen University, 2006

[17] Beasley M S.An Empirical Analysis of the Relation between the Board of Director Composition and Financial Statement Fraud[J].*The Accounting Review,1996,71(4):443-465.*

[18] Abbott L.J., Parker S., Peters G.F. Audit Committee Characteristics and Financial Misstatement[J]. *Auditing: A Journal of Practice&Theory,2002,23(1):69-87.*

[19] Beasley M S,Carcello J.V.,Hermanson D.R.et al. Fraudulent Financial Reporting: Consideration of Industry Traits and Corporate Governance Mechanisms[J].*Accounting Horizons,2000,14(4):441-454.*

[20] LIguo Liu, Ying Du. Empirical Study on the relationship between Company governance and accounting information quality[J]. Accounting Study 2003, （2） : 28-36

[21] Fanning K. and Cogger K., (1998), 'Neural Network Detection of Management Fraud Using Published Financial Data', International Journal of Intelligent Systems in Account-ing, Finance & Management,Vol. 7, No. 1, pp. 21-24.

[22] Lin J W,Hwang M I,Becker J D. A Fuzzy Neural Networks for Assessing the Risk of Fraudulent Financial Reporting[J].*Managerial Auditing Journal,2003,18(8):657-665.*

[23] Kirkos E,Spathis C,Manolopoulos Y. Data Mining Techniques for the Detection of Fraudulent Financial Statements[J].*Expert Systems with Applications,2007,32:995-1003.*

[24] Hansen,J.V.,J.B. McDonald,W.F.Messier,Jr.,and T.B.Bell , A generalized qualitative-response model and the analysis of management fraud, 1996, *Management Science,42 : 1022-1032.*

[25] Beneish M.D. Detecting GAAP Violation: Implications for Assessing Earnings Management among Firms with Extreme Financial Performance [J].Journal of Accounting and Public Policy,1997,16:271-309.

[26] Eining M.M.,Jones D.R.and Loebbecke J.K. Reliance on Decision Aids: An Examination of Auditors'Assemment of Management Fraud[J]. *Auditing: A Journal of Practice and Theory,1997(16):1-19*

[27] Guoxin Chen, Zhanjia lv, Feng He. Empirical Study on detecting financial statements Fraud- based on empirical data of public companies. [J]Auditing Study, 2007（3）

[28] Haisong Ren. Investigation Research on Public company financial report fraud. [D] Dalian: Dongbei University of Finance, 2006

[29] Dianmin Yue. Research on model feature and detecting of accounting fraud of China public companies. [D] . Tianjin: Tianjin University of Finance, 2008