Expert Systems With Applications 62 (2016) 32–43

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning

Yeonkook J. Kim[a], Bok Baik[b], Sungzoon Cho[c,*]

[a] Technology Management, Economics and Policy Program, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-744, Republic of Korea
[b] College of Business Administration, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-744, Republic of Korea
[c] Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-744, Republic of Korea

ARTICLE INFO

Article history:
Received 3 February 2016
Revised 12 April 2016
Accepted 9 June 2016
Available online 11 June 2016

Keywords:
Financial misstatement detection
Financial restatements
Fraud intention
Multi-class cost sensitive learning

ABSTRACT

We develop multi-class financial misstatement detection models to detect misstatements with fraud intention. Hennes, Leone and Miller (2008) conducted a post-event analysis of financial restatements and classified restatements as intentional or unintentional. Using their results (along with non-misstated firms) in the form of a three-class target variable, we develop three multi-class classifiers, multinomial logistic regression, support vector machine, and Bayesian networks, as predictive tools to detect and classify misstatements according to the presence of fraud intention. To deal with class imbalance and asymmetric misclassification costs, we undertake cost-sensitive learning using MetaCost. We evaluate features from previous studies of detecting fraudulent intention and material misstatements. Features such as the short interest ratio and the firm-efficiency measure show discriminatory potential. The yearly and quarterly context-based feature set created further improves the performance of the classifiers.

1. Introduction

Can we detect accounting fraud? How are intentional financial misstatements different from accounting irregularities without managerial intent? Answering these questions is of critical importance to the efficient functioning of capital markets and to increase our understanding of financial statement fraud. Fraudulent financial statements affect not just shareholders, but also lenders, creditors and employees. Perols (2011) estimates the cost of financial statement fraud in the U.S. to be $572 billion per year.

Due to the significance of this topic, academics have performed post-event studies extensively to understand the causes, motivations, and consequences of financial misstatements and earnings manipulation (Beneish, 1999; Dechow, Ge, & Schrand, 2010; Dechow, Sloan, & Sweeney, 1995, 1996; DeFond & Jiambalvo, 1994; Ettredge, Scholz, Smith, & Sun, 2010; Gillett & Uddin, 2005; Hennes, Leone, & Miller, 2013; Jones, Krishnan, & Melendrez, 2008; Palmrose & Scholz, 2004; Schrand & Zechman, 2012). Building upon these studies, various prediction/detection models have been proposed in the accounting and data mining literature (Abbasi, Albrecht, Vance, & Hansen, 2012; Beneish, 1999; Cecchini, Aytug, Koehler, & Pathak, 2010; Dechow, Ge, Larson, & Sloan, 2011; Huang, Tsaih, & Yu, 2014; Kirkos, Spathis, & Manolopoulos, 2007; Kotsiantis, Koumanakos, Tzelepis, & Tampakas, 2006; Lin, Chiu, Huang, & Yen, 2015; Pai, Hsu, & Wang, 2011; Ragothaman, Carpenter, & Buttars, 1995).

To develop a detection/prediction model, databases which contain financial restatements, securities class action lawsuits, and financial regulatory investigation data, such as U.S. Securities and Exchange Commission (SEC) enforcement samples, have been used as a binary target variable. However, financial misstatements can be classified as involving either errors (i.e., unintentional misapplications of accounting rules) or irregularities (i.e., intentional misreporting). Previous studies often assume intentional misreporting either explicitly or implicitly, but the samples do contain both errors and irregularities (Beasley, 1996; Hennes, Leone, & Miller, 2008).

An intentional misstatement arises when management has incentives to manage earnings or commit fraud to meet certain objectives, such as maximizing personal gain through stock-based compensation (Erickson, Hanlon, & Maydew, 2006). When a restatement, the revision and publication of one or more of a company's previous financial statements with a material inaccuracy, is announced, investors show different responses depending on the presence of fraud intention. For example, Palmrose, Richardson, and Scholz (2004) report that the market reacts to restatement announcements differently showing an average abnormal return of −20% for financial restatements caused by deliberate

* Corresponding author. Fax: +82 2 889 8560.
E-mail addresses: yjkim@dm.snu.ac.kr (Y.J. Kim), bbaik@snu.ac.kr (B. Baik), zoon@snu.ac.kr (S. Cho).

http://dx.doi.org/10.1016/j.eswa.2016.06.016
0957-4174/

misreporting as opposed to an average abnormal return of −6% for non-fraud restatements. Also, when intentional misstatements are announced, a higher CEO/CFO turnover rate and more frequent securities class-action lawsuits follow compared to when unintentional restatements are announced (Hennes et al., 2008).

In comparison to intentional misstatements, unintentional misstatements are more likely to result from weak internal controls, to occur more frequently, to affect a broader range of accounts, and are likely to lead to auditor turnover (Hayes, 2014). Moreover, unintentional errors indicate either management is less reputed (Demerjian, Lev, Lewis, & McVay, 2012) or less incentivized to implement and maintain effective controls over financial reporting (Hayes, 2014).

Due to the difference between intentional and unintentional misstatements, researchers testing hypotheses involving managerial misconduct are at risk of making incorrect inferences regarding their hypotheses if they do not specifically distinguish intentional misstatements from unintentional errors (Hennes et al., 2008). This is especially critical given that the relative frequency of error-related misstatements has increased due to the tighter regulation in the post-Enron regulatory environment. Furthermore, if researchers limit their samples to only fraudulent misstatements for their detection models, they are underutilizing information by throwing away more commonly occurring unintentional misstatements. As a result, their models may not effectively detect more frequent but less egregious misstatements or discriminate between intentional misstatements and unintentional errors. We fill the void in the literature by distinguishing intentional financial misstatements from unintentional misstatements and by presenting a fraud-detection model.

To the best of our knowledge, our study is the first predictive study that classifies financial misstatements according to the presence of fraud intention using multi-class classifiers. Specifically, in this study, we investigate what causes intentional and material misstatements by classifying instances into three groups: 1. Intentional misstatement (Irregularity); 2. Unintentional misstatement (Error); and 3. No misstatement. To deal with asymmetric misclassification costs, we undertake cost-sensitive learning using MetaCost.

The contributions of this paper go further than filling a void in the literature by developing the first multi-class predictive models alone. Our study provides a quantitative tool to detect fraud intention of senior management of public firms. This should benefit academics and practitioners in financial regulation and capital markets. More specifically, regulators such as the SEC would benefit from our work because they could focus their investigation efforts on cases that are more likely to involve fraudulent intention. Also, investors and financial institutions would benefit from appropriately adjusting their levels of exposure to suspected firms in advance. Moreover, auditors can tailor their audit processes accordingly and minimize their possible legal risks.

In the following section, we review relevant literature. In Section 3, we describe the data used and the testing methodology employed. We then present the results in Section 4. Section 5 concludes this paper.

2. Financial misstatement literature

2.1. Fraud detection

West and Bhattacharya (2016) group common types of financial fraud into three groups: bank fraud (e.g. credit card fraud, Mortgage fraud, money laundering), corporate fraud (e.g. financial statement fraud, securities and commodities fraud) and insurance fraud (e.g. automobile insurance fraud, health card fraud). Researchers have analyzed various types of financial fraud (Bolton & Hand, 2002; Fawcett & Provost, 1997) and have proposed statistical and machine learning methods to detect fraud effectively (Ngai, Hu, Wong, Chen, & Sun, 2011; West & Bhattacharya, 2016). For example, Dal Pozzolo, Caelen, Le Borgne, Waterschoot, and Bontempi (2014) and Van Vlasselaer et al. (2015) propose credit card fraud detection models.

Among various types of financial fraud, Accounting researchers have performed post-event studies extensively to understand the causes, motivations, and consequences of financial statement fraud (Badertscher, 2011; Beasley, 1996; Beneish, 1999; Dechow et al., 1995, 1996; Erickson et al., 2006; Jones et al., 2008). As shown in Table 1, many of these post-event studies use the binary target variable made of fraud (misstatement) firms and non-fraud (non-misstatement) firms.

2.2. Classification of material misstatements according to fraudulent intention (post-event studies)

Researchers recently began to investigate and classify financial material misstatements according to management intent to mislead, manipulate or defraud, rather than to simply tag them all as examples of fraud. For example, Beasley (1996) searches Accounting and Auditing Enforcement Releases (AAERs) which as issued by the SEC during or at the conclusion of an investigation against a company, an auditor, or an officer for alleged accounting and/or auditing misconduct. He identifies fraud firms by removing AAERs not involving financial statement fraud (e.g., unintentional misapplication of GAAP).

In an award winning study, Hennes et al. (2008) formally propose the following three rules to classify financial restatements as errors (unintentional) or irregularities(intentional):

1. Classify any restatements using variants of the words “fraud” or “irregularity” in reference to the misstatement in 8-K filings as irregularities
2. Classify restatements with related SEC or Department of Justice investigations as irregularities
3. Presence or absence of other investigations into accounting matter (e.g., the audit committee hires a forensic accounting firm): classify restatements with related independent investigations as irregularities.

They perform three validity tests to support their classification approach. The first validity test shows a significant difference in the stock market reactions between the two groups: the mean (median) cumulative abnormal return for the unintentional-misstatement sample was 1.93%(0.90%) compared to 13.64%(19.4%) for the intentional-misstatement sample. The second validity test compares the frequency of securities class-action lawsuits, showing that 84 of the 105 intentional misstatements in their sample had contemporaneous class-action lawsuits while one of the 83 unintentional-misstatement samples had a related lawsuit. The third validity test shows that the percentage of restating firms experiencing CFO/CEO turnover in the 13 months surrounding the restatements (six months before to six months after) was 49%(64%) for CEOs(CFOs) in their intentional-misstatement sample but only 8%(12%) in the unintentional-misstatement sample. In the analysis for CFO/CEO turnover, they showed that the power of the hypotheses test on accounting restatements significantly improved either by limiting restatement samples to intentional misstatements or by including a control variable distinguishing unintentional misstatements from intentional misstatements. In a later study, Plumlee and Yohn (2010) classify financial restatements into four groups: intentional manipulation, internal company error, transaction complexity, and accounting standards.

Hayes (2014) proposes a simple text-search approach to classify financial restatements as unintentional errors or intentional

Table 1
Summary of literature review.

	Binary target variable	Tertiary target variable
Post-event analyses	Badertscher (2011); Beasley (1996); Beneish (1999); Dechow et al. (1995, 1996); Erickson et al. (2006); Jones et al. (2008)	Hayes (2014); Hennes et al. (2008); Plumlee and Yohn (2010)
Predictive studies	Abbasi et al. (2012); Beneish (1999); Cecchini et al. (2010); Dechow et al. (2011); Green and Choi (1997); Kirkos et al. (2007); Kotsiantis et al. (2006); Pai et al. (2011); Perols (2011)	Our work

misstatements, or as unclassified. She analyzes how intentional misstatements and unintentional errors differ, finding first that unintentional errors are associated with a market reaction of a smaller magnitude, are more likely to be associated with weak internal controls, occur more frequently, affect a broader range of accounts, and are likely to lead to auditor turnover if a restatement occurs (Hennes et al., 2008; Palmrose & Scholz, 2004; Plumlee & Yohn, 2010). Moreover, whereas external users and audit committee members can rationally infer the presence of intentional misstatements by managerial incentives to manage earnings or commit fraud to meet certain targets, unintentional errors are less transparent and thus may be more likely to result in decision errors.

Second, unintentional errors may reflect the (lack of) competence or ability rather than the integrity of management and auditors. While intentional misstatements may be partially due to management incentives, unintentional errors can indicate either managements inability (Demerjian et al., 2012) or a lack of incentive to implement and maintain effective controls over financial reporting (Hoitash, Hoitash, & Johnstone, 2012).

Third, management may react differently to an auditors detection of intentional vs. unintentional misapplications of GAAP. Management motivated by concerns such as shareholder reactions to missing analysts forecasts is more likely to resist correcting an intentional misstatement. Non-fraud-intended management, on the other hand, should be less resistant to correcting material unintentional errors. Therefore, restatements that correct unintentional errors indicate that management has failed to implement and maintain effective controls over financial reporting and that the auditor lacks the requisite skills to plan and conduct an effective audit.

In short, these studies show that intentional misstatements and unintentional misstatements are two different types of events and that distinction between the two is important to increase the power of the tests. Moreover, these studies are *ex-post* studies, suggesting that *ex-ante* studies are warranted. According to a report by the Committee of Sponsoring Organizations of the Treadway Commission on the SEC investigation between 1998 and 2007, the median fraud period was two years. This implies that it takes approximately two years for financial statement fraud to be identified, investigated and announced. For example, on November 8th, 2001, Enron Corp. announced that it would restate earnings for the period 1997–2001. Then, on December 2nd, 2001, Enron and 13 of its subsidiaries filed for Chapter 11 bankruptcy protection (GAO, 2002). This suggests that financial fraud was revealed approximately three years after its first misstated financial statements had been issued. Our goal is to detect Enrons misstatement in the first year it takes place and predict if Enron intended to defraud and thus mislead readers of its financial statement.

2.3. Binary prediction/Detection models

Most financial misstatement detection/prediction models deal with a binary problem. For illustration purposes, we summarize the classification result of Hennes et al. (2008) in Fig. 1. Studies either classify intentional-misstatement firms (or a subset of A as fraudulent firms[1]) and non-misstatement firms (i.e. A vs. D, e.g. Cecchini et al. (2010)) or classify misstatement data and non-misstatement firms (C vs. D, where C=A+B, e.g. Dechow et al. (2011)).

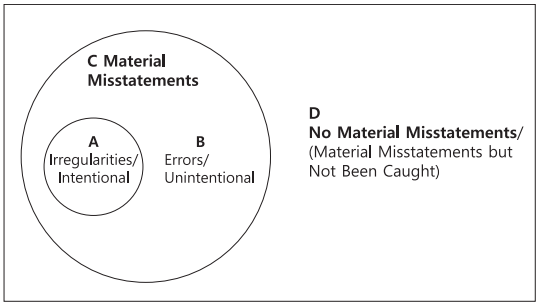

Fig. 1. Hennes et al. (2008) classification result.

We describe several previous studies. First, Beneish (1999) classifies intentional earnings manipulators and non-manipulators (A vs D) and develops a probit model termed the M-score model using eight financial ratios to predict cases of upward earnings manipulation. The study explains that the model may make two types of errors; it can classify a company as a non-manipulator when it manipulates (a Type I error), or it can classify a company as a manipulator when it does not manipulate (a Type II error). The probability cutoffs that minimize the expected costs of misclassification depend on costs associated with the relative cost of making an error of either type. He compares the expected costs by increasing relative costs of Type I to Type II errors from 1:1 to 100:1.

Extending the study of Beneish (1999), Dechow et al. (2011) classify misstated and non-misstated firms (C vs. D) and develop F-score models using logit models. Unlike the M-score model, which uses only financial statement variables, the F-score models use financial statement variables, market-related variables, off-balance sheets and other nonfinancial variables.

Applying a more flexible modelling approach, Cecchini et al. (2010) classify intentionally misstated (fraudulent) and non-misstated firms (A vs. D) using a support vector machine classifier that incorporates a custom financial kernel. The financial kernel is a graph kernel that uses input financial variables to derive implicitly numerous financial ratios (i.e., 24 financial statement variables into 1518 features in the study). They argue that because fraud tactics change over the years, a method that utilized exhaustive combinations of potential fraud variables has a better chance of effectively catching fraud as compared to methods that restrict themselves to a few possible constructs. They show that their model outperforms a logit model (F-score model) and the neural network model by Green and Choi (1997) and argue that the superior

[1] Although the distinction between fraud and irregularities has become blurred over the years, the two terms are technically not identical (Hennes et al., 2008). Studies such as that by Erickson et al. (2006) make a strict distinction between the two. However, auditing guidelines (e.g. SAS No. 82, AICPA 1997) use the term "fraud" and the term "intentional misstatements" interchangeably (Hennes et al., 2008).

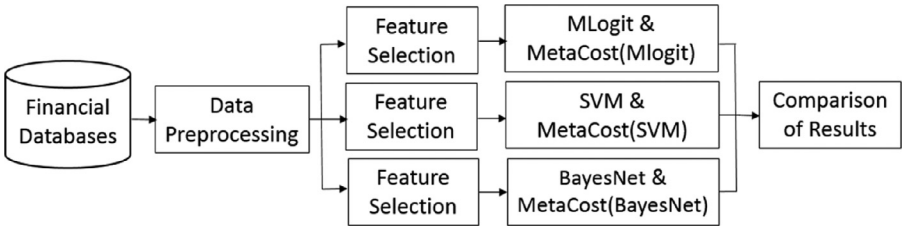

Fig. 2. Research framework.

performance stems more from the additional features used than from the specific induction technique used. To counter class imbalance and non-symmetric misclassification costs, they apply different weightings for fraud and non-fraud classes and achieve their best result at 200:1.

Abbasi et al. (2012) also classify fraudulent firms and non-fraud firms (A vs D) using ensemble and adaptive learning. Similar to Cecchini et al. (2010), the authors argue that prior financial fraud-detection studies utilize feature sets that are too small in number and exclusively from annual statements, leading to feature sets lacking representational richness and that are simply not large enough to generate appropriate hypothesis spaces for the classification methods utilized. To alleviate this problem, they incorporate the following refinements: (1) the inclusion of organizational and industry-level context information and (2) the utilization of data based on quarterly and annual statements.

The study uses twelve financial statement ratios as seed variables and constructs, first, the organizational context features by computing the difference between (-) and the ratio of (/) the firms current period seed financial ratios and previous time period ratios. Second, they construct the industry-level context features using two industry-representative models; Top-5 (industry leaders) models are created by averaging the data from the five largest companies in each industry-year (in terms of sales), creating what Whiting, Hansen, McDonald, Albrecht, and Albrecht (2012) refer to as "centroid" firms. The twelve seed financial ratios from centroid firms are then generated and compared to misstated/non-misstated firm ratios. Whiting et al. (2012) argue that industry-representative centroid models measure a majority influence on what is normal for each industry. Also, by averaging the top five firms in each industry, the individual differences between companies are smoothed, providing a better industry-representative model. Furthermore, given that single comparison companies cannot be guaranteed not to be fraudulent or misstated, this method provides a more robust and theory-driven context.

Similarly, closest-5 models (industry peers) are created for each firm by averaging the data from the five companies in the same industry-year that are most similar in terms of sales. Organizational context features and industry-level context features are created using both annual data and quarterly data. Therefore, 84 yearly context-based feature sets and 336 quarterly context-based feature sets are created and used to train 14 base classifiers. They perform stacking and semi-supervised learning and achieved a fraud-detection rate of 88% and an area-under-the-receiver operating characteristic curve which exceeded 0.9.

In summary, two-class models that classify C and D (as in Fig. 1) measure the patterns of both intentional misstatements and unintentional misstatements. On the other hand, binary models that classify A and D only measure the difference between intentional misstatements and non-misstatements. Unless the patterns of intentional misstatements and unintentional misstatements are ordinal in the measurements, this approach may not be effective when used to detect material but unintentional misstatements.

3. Research methodology

As shown in Fig. 2, our proposed research framework consists of several steps. First, we gather restatement and financial data from various databases such as Compustat, a database of financial, statistical and market information. Second, in the data preprocessing step, we create variables/ratios based on the past accounting and datamining studies. Regarding missing values and outliers, we follow the guidelines in the literature that use the variables. For example, the Asset Quality index variable (the first variable in Table 2) that measures year-to-year changes, we winsorize the data at the 1% and 99% percentiles. Then, we set the missing values to 1 as described in Beneish (1999). Third, we perform feature selection process for each model separately. Using Weka 3.7.12, a popular machine learning software, we test both the filter method (e.g. Info gain attribute evaluator) and the wrapper method (e.g. Bayesian Network model with the genetic algorithm search). We select the feature selection process that gives the best performance for each classifier. For example, for the Bayesian Network model, we use the wrapper method with the Particle Swarm Optimization algorithm to do feature selection. Lastly, we build the models and compare the result.

3.1. Variables

We use the post-event classification data on financial restatements from Hennes et al. (2008) as a three-class target variable. Our misstated dataset contains 788 instances with 214 instances of irregularities and 355 as errors for the period of 1992 to 2005. The non-misstated dataset contains 2156 instances selected based on industry and fiscal year.

We select and test various features from prior post-event studies and detection studies. Financial statement fraud evolves with time as perpetrators find new and clever ways to circumvent implemented procedures and controls. At the same time, new regulations or weak internal control systems in rapidly growing firms may cause unintentional misstatements. In order to effectively detect and discriminate between intentional misstatements and unintentional errors, it is necessary to use not just financial statement data but also various other features that are found to have some predictability in detecting financial misstatements. Researchers have found that off-balance sheet variables (e.g., operating leases), nonfinancial measures (e.g., abnormal changes in employees), market variables (e.g. market-adjusted stock returns) and governance measures (e.g. CEO power) may signal the presence of possible conditions of accounting manipulation. We list the final selection of 49 features included in our experiment and the corresponding references in Table 2. The Y/Q column indicates whether a feature is an annual ratio/measure or a quarterly ratio/measure.

Following Abbasi et al. (2012), we create top-5 models (industry leaders) and closest-5 models (industry peers) for the variables that have sufficient data. For a yearly financial ratio, for an example, we compare it with the top-five firm (sales-weighted) average value and with the average value from five peer firms. In addition,

Table 2
Feature list.

No	Variable	Definition	Y/Q	Reference
1	Asset quality index	Ratio of non-current assets other than property, plant, and equipment (PP&E) to Total Assets (TA), for time period t relative to time period t-1	Y, Q	Abbasi et al. (2012); Beneish (1999)
2	Inventory growth	Inventory at period t/ Inventory at period t-1	Y, Q	Abbasi et al. (2012)
3	Receivable to assets	Accounts Receivable(AR) /TA	Y, Q	Dechow et al. (2011)
4	Soft assets	(TA -PP&E - Cash & Cash Equivalent)/TA	Y, Q	Dechow et al. (2011)
5	PP&E to TA	PP&E / TA	Y, Q	Perols (2011)
6	Receivable	Accounts Receivable	Y, Q	Perols (2011); Green and Choi (1997)
7	RSST accrual	Accrual measure following Richardson, Sloan, Soliman, and Tuna (2005)	Y	Dechow et al. (2011)
8	Accruals to assets	Change in working capital accounts other than cash less depreciation/TA	Q	Beneish (1999)
9	Cash flow earnings correlation	Correlation between earnings and cash flow from operation	Q	Dichev, Graham, Harvey, and Rajgopal (2013)
10	Earnings smoothness	Standard deviation of earnings divided by standard deviation of cash flow from operation	Q	Dichev et al. (2013)
11	Asset turnover	Net sales/Total assets	Y, Q	Abbasi et al. (2012);
12	Operating margin	Net income/Net sales	Y, Q	Abbasi et al. (2012)
13	Depreciation index	Ratio of the rate of depreciation in period t-1 to the corresponding measure in period t	Y, Q	Abbasi et al. (2012); Beneish (1999)
14	Days sales in receivables	Ratio of day sales in receivables in period t to the corresponding measure in period t-1	Y, Q	Abbasi et al. (2012); Beneish (1999)
15	Gross margin index	Ratio of the gross margin in period t-1 to the gross margin in period t	Y, Q	Abbasi et al. (2012); Beneish (1999)
16	SGE expense	Ratio of selling and general administrative expenses to net sales in period t by the same ratio in period t-1	Y, Q	Abbasi et al. (2012); Beneish (1999)
17	Sales Growth	Net sales in period t divided by net sales in period t-1	Y, Q	Abbasi et al. (2012); Beneish (1999)
18	Change in Cash sales	Percentage change in cash sales where cash sales = (Sales - ΔAccounts Receivable)	Y, Q	Dechow et al. (2011)
19	Change in return on assets	Earnings/Average TA in period t divided by the same ratio in period t-1	Y, Q	Dechow et al. (2011)
20	Bloat	Net operating assets in year t divided by total sales at the end of year t-1	Y, Q	Ettredge et al. (2010); Badertscher (2011)
21	Inverse of the firms interest coverage ratio	Interest expense in year t, divided by operating income before depreciation in year t-1;	Y, Q	Badertscher (2011)
22	Special Items as a Percentage of Sales	Special Items / Sales	Y, Q	McVay (2006)
23	Deferred tax expense	Deferred tax expense for period t âüä total assets for t-1	Y, Q	Dechow et al. (2011)
24	Debt to Equity Ratio	Debt/Equity	Y, Q	Perols (2011)
25	Firm efficiency	Revenue generating ability given level of resources such fixed assets and R&D expenses	Y	Demerjian et al. (2012, 2012)
26	Unexplained Discretionary expenses	Error term from industry discretionary expense regression equation	Y	Roychowdhury (2006)
27	Unexplained Production costs	Error term from industry production cost regression equation	Y	Roychowdhury (2006)
28	Cash flow from operations	Cash flow from operations at period t scaled by asset at period t-1	Y	Roychowdhury (2006)
29	Unearned Revenue Long Term over Revenue	Long-term Deferred Revenue/ Revenue	Y, Q	GMI Ratings (2013)
30	Accounts Payable over Operating Expenses	Accounts Payable / Operating Expenses	Y, Q	GMI Ratings (2013)
31	Abnormal change in employees	Percentage change in the number of employees /percentage change in assets	Y	Dechow et al. (2011)
32	Change in operating lease activity	Change in the present value of future noncancelable operating lease obligations deflated by average total assets	Y	Dechow et al. (2011)
33	Existence of operating leases	Indicator variable coded 1 if future operating lease obligations are greater than zero	Y	Dechow et al. (2011)
34	Book-to-market ratio	(Asset-Liabilities)/Market capitalization	Y	Badertscher (2011); Ettredge et al. (2010)
35	Market value	Natural log of the market value of equity	Y	Ettredge et al. (2010)
36	Actual issuance	Indicator variable coded 1 if the firm issued securities during year t	Y	Dechow et al. (2011)
37	Acquisition	Indicator variable equal to 1 if firm had an acquisition that contributed to sales in the prior year is greater than 0, zero otherwise	Y	Ettredge et al. (2010)
38	Meeting Analyst forecast	Indicator variable equal to 1 if the firm meets or beats the median annual analyst earnings forecast; 0 otherwise	Y	Badertscher (2011); Perols (2011)
39	Analyst forecast error	Median analyst earnings forecast - actual EPS	Y	Badertscher (2011)
40	CEO Bonus	Ratio of the CEOs bonus divided by the total compensation received by the CEO in year t	Y	Badertscher (2011)

(*continued on next page*)

Table 2 (*continued*)

No	Variable	Definition	Y/Q	Reference
41	CEO Chairman	indicator variable that equals 1 if the CEO also holds the title of chairperson in year t, and 0 otherwise	Y	Badertscher (2011)
42	Owner	Sum of restricted stock grants in the current period and the aggregate number of shares held by the executive (excluding stock options) scaled by total outstanding shares	Y	Badertscher (2011)
43	CEO Salary	Amount of CEOs base salary in year t;	Y	Badertscher (2011)
44	Stock options	Number of unexercised options (including options grants in the current period) that the executive held at year-end t-1 scaled by total outstanding shares of the firm	Y	Badertscher (2011)
45	Short interest	Short interest ratio at year end	Y	Dechow et al. (2011)
46	Cash flow from financial activities	Level of finance raised / Average total assets	Y, Q	Dechow et al. (2011)
47	Change in leverage Ratio	Ratio of sum of short-term debt and long-term debt in period t, scaled by total assets in period t-1 divided by the same ratio one period later	Y, Q	Abbasi et al. (2012);
48	Market-adjusted stock return	Annual buy-and-hold return inclusive of delisting returns minus the annual buy-and-hold value-weighted market return	Y	Dechow et al. (2011)
49	Lagged market-adjusted stock return	Previous years annual buy-and-hold return inclusive of delisting returns minus the annual buy-and-hold value-weighted market return	Y	Dechow et al. (2011)

Table 3
Yearly context-based feature set type.

Type	Description
Yearly financial ratio	R1
Industry-level context: Top-5 Model	R1-T1, R1/T1
Industry-level context: Closest-5 Model	R1-C1, R1/C1
Organizational context	R1-P1, R1/P1

we calculate the changes by dividing (or subtracting) the current year ratio by the prior year ratio (as in Table 3), creating six additional variables to be used in our experiments. We repeat the process with quarterly variables. In Table 4, R1Q1 signifies the quarterly ratio number (R1 means the ratio number one) and the number of quarter (Q1 means the first quarter). Since there are four quarters in a given fiscal year, we have four quarterly ratios. For the top-5 model, we compute the difference between (-) and the ratio of (/) the firms current quarterly ratios and the same ratio of the average of five largest companies in the same period. We repeat the same procedure with the averages of five industry peers (firms with the most similar sales amounts in the same industry) to create the closest-5 model variables. The organizational context variables are created by computing the changes of the firms current quarterly ratios from previous quarterly ratios. We generate in total 1,086 features.

3.2. Methods

To perform the classification, we employ three multi-class classifiers: Multinomial logistic regression, Support Vector Machines, and Bayesian Networks. We select the three models based on prior studies. According to Abbasi et al. (2012), various statistical and machine learning models are used in developing financial misstatement detection models. However, the most popular model is the logistic regression model: 10 out of 15 financial statement fraud detection studies in their review employ logistic regression models. Logistic regression models can be used in binary and multi-class classifications. Also, logistic regression models are often used in related accounting literatures because the models are interpretable and statistical inference can be easily made (Beneish, 1999; Dechow et al., 2011; Dechow et al., 1995; Ettredge et al., 2010; Jones et al., 2008).

On the other hand, in more recent predictive studies, support vector machines are a more popular choice due to their high classification accuracy (Abbasi et al., 2012; Cecchini et al., 2010; Pai et al., 2011; Perols, 2011). For example, Perols (2011) compares the performance of six popular statistical and machine learning models in detecting financial statement fraud under different assumptions of misclassification costs and ratios of fraud firms to non-fraud firms. He shows that logistic regression and support vector machines outperform an artificial neural network, bagging, C4.5, and stacking.

Kirkos et al. (2007) also investigate the usefulness of decision trees, neural networks and Bayesian networks in the fraudulent financial statement detection and show that the Bayesian network model achieves the best result. Bayesian network models can do multi-class classification, directly produce class probability estimates (which is useful for the cost-sensitive learning later), and performs well even when the data is highly class-imbalanced (Leong, 2015).

3.2.1. Multinomial logistic regression (MLogit)

Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems. It predicts the probabilities of different possible outcomes of a categorically distributed target variable given a set of independent features. More specifically, it models the posterior class probabilities Pr(G = j|X = x) for J classes via linear functions in x while at the same time ensuring that they sum to one and remain between 0 and 1. The model has the form:

$$Pr(G = j|X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{j} e^{F_k(x)}}, \sum_{k=1}^{j} F_k(x) = 0 \quad (1)$$

where $F_j(x) = \beta_j^T * x$ are linear regression functions. The model is usually fit by finding maximum likelihood estimates for the parameters β_j. Among the variants of logistic regression models, we use SimpleLogistic classifier implemented in Weka 3.7.12, which uses the LogitBoost algorithm to fit the logistic models. LogitBoost

Table 4
Quarterly context-based feature set type.

Type	Description
Quarterly financial ratios	R1Q1, R1Q2, R1Q3, R1Q4
Industry-level context: Top-5 Model	R1Q1-T1Q1, R1Q2-T1Q2, R1Q3-T1Q3, R1Q4-T1Q4 R1Q1/T1Q1, R1Q2/T1Q2, R1Q3/T1Q3, R1Q4/T1Q4
Industry-level context: Closest-5 Model	R1Q1-C1Q1, R1Q2-C1Q2, R1Q3-C1Q3, R1Q4-C1Q4 R1Q1/C1Q1, R1Q2/C1Q2, R1Q3/C1Q3, R1Q4/C1Q4
Organizational context	R1Q2-R1Q1, R1Q3-R1Q2, R1Q4-R1Q3, R1Q1-R1Q4 R1Q2/R1Q1, R1Q3/R1Q2, R1Q4/R1Q3, R1Q1/R1Q4

performs the forward stage-wise fitting of additive logistic regression models by generalizing the Eq. (1) to $F_j(x) = \Sigma_m f_{mj}(x)$, where f_{mj} can be arbitrary functions of the input variables that are fit by least squares regression. The SimpleLogistic classifier determines the best number of LogitBoost iterations by cross-validation. During the cross-validation process, only those attributes that improve the performance are included. In this manner, automatic feature selection is performed (Landwehr, Hall, & Frank, 2005).

3.2.2. Support vector machines (SVM)

SVM use a linear model to implement nonlinear class boundaries by mapping input vectors nonlinearly into a high-dimensional feature space. In the new space, an optimal separating hyperplane is constructed. The training examples that are closest to the maximum margin hyperplane are known as support vectors. All other training examples are irrelevant with regard to defining the binary class boundaries. Good separation is achieved by the hyperplane with the greatest distance to the nearest training-data point of any class, as in general the larger the margin, the lower the generalization error of the classifier (Han, Kamber, & Pei, 2011).

SVM is directly applicable for two-class tasks. Therefore, it is necessary to apply algorithms that reduce a multi-class task to several binary problems. The most common approach is to build binary classifiers using one of the classes and the rest (one-versus-all) or with every pair of classes (one-versus-one).

Unlike a probabilistic classifier (such as logistic regression) that is able to predict, given a sample input, a probability distribution over a set of classes, SVM produces only scores as the output. To obtain proper probability estimates, a common approach in the binary case is to apply Platt scaling, which learns a logistic regression model with the scores (Platt, 1999). In the multi-class case, the predicted probabilities can be coupled using the pairwise coupling algorithm by Hastie and Tibshirani (Hastie, Tibshirani et al., 1998; Witten & Frank, 2005).

3.2.3. Bayesian networks (BayesNet)

Bayesian classification is based on the Bayes theorem, a method to determine the probability that a given hypothesis is true. The theorem states that for a hypothesis H (such as whether an object X can be classified within a given class), the probability P is given as follows:

$$P(H|X) = \frac{P(X|H) * P(H)}{P(H)} \quad (2)$$

A Bayesian classifier calculates P(C_i|X) for all possible classes and inserts X into the class with the highest conditional probabilities.

Bayesian networks are probabilistic graphical models that allow the representation of dependencies among subsets of attributes. More specifically, a Bayesian Network is a directed acyclic graph, where each node represents an attribute and each arrow represents an instance of probabilistic dependence. If an arrow is drawn from node A to node B, then A is the parent of B and B is a descendent of A. In a Bayesian Network, each variable is conditionally independent of its nondescendents, given its parents (Han et al., 2011). For each node X, there exists a conditional probability table, which specifies the conditional probability of each value of X for each possible combination of the values of its parents. The probability of an instance having m attributes is expressed as shown in Eq. (3) below.

$$P(X_1, X_2, ..., X_m) = \prod P(X_i|Parents(X_i)) \quad (3)$$

The network structure can be defined in advance or can be inferred from the data. To perform the classification, one of the nodes can be defined as the class node, with the network then calculating the probability of each alternative class.

Table 5
Two-class cost matrix (2x2).

	Predicted misstated	Predicted not misstated
Actual misstated (= 0)	C(0,0) or TN	C(1,0) or FN
Actual not misstated (= 1)	C(0,1) or FP	C(1,1) or TP

3.2.4. Cost-sensitive learning

Cost-Sensitive Learning is a type of learning in data mining that takes the misclassification costs (and possibly other types of cost) into consideration. Its goal is to minimize the total cost (Ling & Sheng, 2010). Cost-sensitive learning methods, such as the MetaCost procedure, deal with class-imbalance by incurring different costs for different classes (Ling & Sheng, 2010). It is feasible to handle unequal misclassification costs and class-imbalance in a unified framework using cost-sensitive learning as long as the data is not very severely class-imbalanced (Liu & Zhou, 2006). Prior research proposed various cost-sensitive learning methods for multi-class classification tasks (Bermejo, Gámez, & Puerta, 2011; Bourke, Deng, Scott, Schapire, & Vinodchandran, 2008; Domingos, 1999; Sun, Kamel, & Wang, 2006; Witten & Frank, 2005; Xia et al., 2009; Zhou & Liu, 2006, 2010).

Ling and Sheng (2010) categorize cost-sensitive learning into two categories. The first category is to design classifiers that are cost-sensitive in themselves. One example is a cost-sensitive decision tree (Drummond & Holte, 2000). The second category is to design a "wrapper" that converts any existing cost-insensitive classifiers into a cost-sensitive classifier. They further classified the wrapper method into sampling and thresholding categories. Sampling modifies the class distribution of training data based on misclassification costs and then applies cost-insensitive classifiers to the sampled data directly. Weighting can also be viewed as a sampling method. It assigns a normalized weight to each instance according to the misclassification costs. Examples of a rare class with a higher misclassification cost are assigned proportionally higher weights. This can be viewed as example duplication that increases the sample sizes of the rare class.

To explain thresholding, we briefly introduce a theory of cost-sensitive learning using a simple binary example. However, this example can be easily extended to multiple classes. In a binary classification, classification costs can be represented in a cost matrix where two types of correct classification, true positives (TP) and true negatives (TN), and two types of error, false positives (FP) and false negatives (FN), have different costs and benefits. As shown in Table 5, C(i, j) represents the cost of classifying an instance

belonging to class j into class i. Cost(0,0) and Cost(1,1) are usually considered benefits, while the other two cases are costs.

Given the cost matrix, an instance should be classified into a class that generates the minimum expected cost. The expected cost (or conditional risk) of classifying an instance x into class i, R(i|x), can be expressed as

$$R(i|X) = \sum_j P(j|X) * C(i, j), \tag{4}$$

where P(j|x) is the probability of class j being the actual class of instance x (Kim, Choi, Kim, & Suh, 2012). The classifier will then classify instance x into the not misstated class (where class 1 is not misstated) if

$$P(0|X)C(1,0) + P(1|X)C(1,1) \leq P(0|X)C(0,0) + P(1|X)C(0,1) \tag{5}$$

If we rearrange the Eq. (5):

$$P(0|X)(C(1,0) - C(0,0)) \leq P(0|X)(C(0,1) - C(0,0)) \tag{6}$$

If we assume C(0,0) = C(1,1), and because P(0|x) = 1-P(1|x), we can determine the threshold p∗ for the classifier to classify an instance x as non-misstated if (1|x) ≥ p∗, where

$$P^* = \frac{C(1,0)}{C(1,0) + C(0,1)} = \frac{FP}{FP + FN} \tag{7}$$

If a cost-insensitive classifier can produce a posterior probability estimation p(1|x) for test example x, we can make it cost-sensitive by simply choosing the classification threshold using Eq. (7) and classify any example to be non-misstated whenever P(1|x) ≥ p∗. Thresholding uses Eq. (7) as a threshold to classify examples into positive or negative if the cost-insensitive classifiers can produce probability estimations.

The cost-sensitive learning method we employ in this paper is a wrapper method known as MetaCost (Domingos, 1999). MetaCost is a thresholding method. It first uses bagging on the base classifier to obtain reliable probability estimations of training examples and relabels the classes of training examples according to Eq. (7) (Ling & Sheng, 2010). The advantages of using MetaCost are that it is applicable to multi-class classification problems and can even be used with classifiers that do not produce class probabilities directly. In brief, MetaCost works by (1) sampling multiple bootstrap replicates of the training set, (2) learning the classifier of each set, (3) producing the average of the class probabilities that are directly yielded by the classifiers or calculating the fraction of votes received from the ensemble, (4) using expected costs to relabel each training instance, and (5) reapplying the classifier to the relabeled training set. Although class probabilities from the base classifier are not required for MetaCost to work, the use of base classifier class probabilities appears to improve the result.

4. Results

4.1. Feature importance

To assess the impact of features in the classification process, we list features selected by multinomial logistic regressions to compute logistic regression functions when estimating the probability of each class. Following Abbasi et al. (2012), we report the features for the multinomial logistic regression model in Table 6. Under each class in Table 6, the two columns present a description of the features and the coefficient values. In the description columns, the letters R, T, C, Q, and P represent the ratio, top-5 industry model, closest-5 industry model, quarter, and previous year, respectively, while the numbers indicate the ratio or quarter. For example, under the intentional-misstatement class, the first feature, R2, signifies the second feature, Inventory Growth in the feature table in Section 3.1. The ratio number corresponds to the number in the feature table. The second example is the sixth feature under the intentional-misstatements class, R5Q1-R5Q4. R5 signifies the fifth feature, the ratio of property, plants, and equipment to the total assets, and Q1 signifies the first quarter. Thus, R5Q1-R5Q4 signifies the difference between the ratio in the first quarter and the same ratio in the fourth quarter of the previous year.

As shown in Table 6, numerous variables related to accruals quality such as changes in inventory (R2) and receivables (R3) are selected along with industry-level and organizational context-based measures. The table provides insight into how the context-based measures supplement the financial ratios, resulting in enhanced financial fraud-detection capabilities.

One of the features selected for the intentional-misstatement class is the firm-efficiency rank (R25), developed by Demerjian, Lev, and McVay (2012). The firm-efficiency measure attempts to estimate the revenue-generating ability of a firm with a given level of resources, such as fixed assets and R&D expenses, by means of a data envelopment analysis, a nonparametric method for the estimation of production frontiers to measure empirically the productive efficiency of decision-making units (Baik, Chae, Choi, & Farber, 2013). The feature has a positive coefficient of 0.46, showing that intentionally misstating firms tend to show higher income-generating abilities given their resources, though not according to the efficient use of their resources but by accounting manipulations. As shown in Fig. 3, while the no misstatement class is evenly distributed in terms of the proportion of firms in each decile (represented by a line graph), a higher proportion of the intentional-misstatement class is found in the top deciles (0.8 to 1) compared to the unintentional classes.

Another interesting selected feature is the short interest ratio (SIR; R45). Short interest refers to the total number of shares of a particular stock that have been sold short by investors but have not yet been covered or closed out. This indicator is used by both fundamental and technical traders to identify the prevailing sentiment held by the market for a specific stock. The coefficient of the SIR for intentional misstated firms is 2.02, suggesting that the higher the SIR is, the more likely the firm is intentionally misstating. This shows that negative sentiment in the financial market may signal the presence of possible conditions of accounting manipulation in the firm.

4.2. Classification result

We undertake classification with three multi-class classifiers, multinomial logistic regression, support vector machine (with a linear kernel), and Bayesian networks, using stratified ten-fold cross validation. As a sanity check, we test how the classifiers classify Enron Corporation, a notorious example of financial fraud. All three models classify the instance correctly, assigning highest class probability levels to the intentional-misstatement class, as shown in Table 7.

To assess the classification performance, we select evaluation measures from prior studies. For example, Sun et al. (2006) use accuracy and the G-mean to compare the performance of cost-sensitive boosting algorithms in multi-class classification problems with imbalanced class distribution. On the other hand, Zhou and Liu (2006) use the total misclassification costs to evaluate the performance of sampling and threshold-moving methods in training cost-sensitive neural networks for both binary and multi-class classification problems. Since each measure used in the past studies has advantages and disadvantages, we use these three measures as a more balanced set of measures of the classification performance. Moreover, since the first objective in building misstatement detection models is to detect material misstatements, we use a measure

Table 6
Features and coefficients by multinomial logistic regression model.

No	Class: intentional misstatements		Class: unintentional misstatements		Class: no misstatements	
	Descr.	Coeff.	Descr.	Coeff.	Descr.	Coeff.
1	R2	3.45	R3-P3	2.69	R2	−4.96
2	R3-P3	2.9	R2Q4-R2Q3	−1.5	R3-P3	−4.85
3	R4	2.21	R4	0.97	R4	−2.75
4	R36	2.02	R36	0.67	R3Q3-R3Q2	2.06
5	R45	2.02	R3Q4	−0.6	R23-P23	1.55
6	R5Q1-R5Q4	1.85	R28Q3-R28Q2	−0.39	R4-P4	−1.12
7	R4Q2	0.7	R11Q1	0.28	R7	−0.78
8	R3	0.64	R46-P46	0.23	R36	−0.74
9	R37	0.48	R49	0.12	R19	0.59
10	R30Q3/R30Q2	−0.47	R47Q3-T47Q3	0.11	R46	−0.45
11	R25	0.46	R5Q2/R5Q1	0.11	R4Q3	−0.33
12	R46Q1-R46Q4	−0.34	R16Q2-C16Q2	0.11	R4-T4	−0.33
13	R46Q3	0.29	R13Q3/C13Q3	0.1	R46-T46	−0.2
14	R38	−0.18	R47Q1/R47Q4	0.06	R9	0.18
15	R16Q2-C16Q2	−0.15	R7-P7	0.05	R11-C11	0.17
16	R49	−0.07	R7-T7	0.03	R19-T19	0.17
17	R17-C17	0.07	R20Q4-R20Q3	0.01	R5/C5	0.1
18	R13Q2/R13Q1	0.02	R47Q4/R47Q3	0.01	R5/P5	−0.07
19	R13-P13	−0.01	R14Q4-T14Q4	0.01	R32/P32	−0.04
20			R22-C22	−0.01	R30/P30	0.03
21					R31-P31	0.02

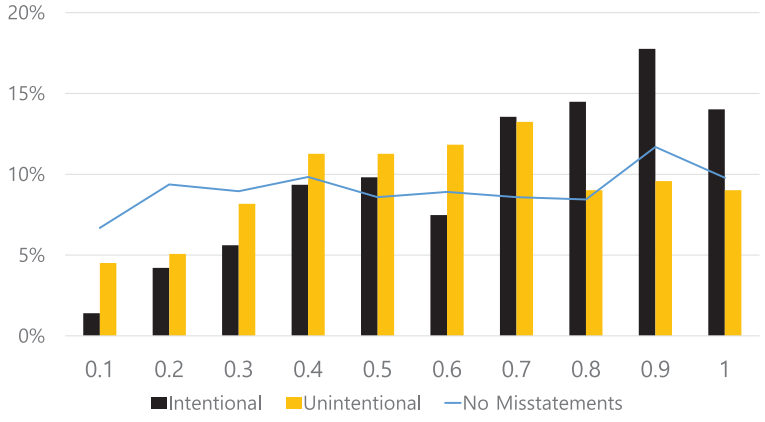

Fig. 3. Firm efficiency deciles for each class.

Table 7
Class probabilities for Enron Inc.

	Intentional	Unintentional	No misstatment
MLogit	0.73	0.13	0.14
SVM-Lin	0.66	0.15	0.20
BayesNet	0.92	0.03	0.05

that shows how well each model detects material misstatements, whether intentional or unintentional, as our fourth measure.

We describe our four evaluation measures.[2] The first measure is accuracy, which is the ratio of the number of instances correctly classified to the total number of instances. The second measure is the G-mean which is calculated as follows,

$$\text{G-mean} = \left(\prod_{i=1}^{k} R_i \right)^{1/k} \tag{8}$$

where R_i is a recall value for class i. As each recall value representing the classification performance of a specific class, G-mean measures the balanced performance among multiple classes of a classification output (Sun et al., 2006). However, this measure does not reflect non-symmetric misclassification costs.

In order to address the issues of non-symmetric misclassification costs in our performance evaluation, we use two other measures. The third measure is the percentage of misstatements detected. Because it would be more costly to misclassify an intentional misstatement as a non-misstatement than to misclassify an intentional misstatement as an unintentional misstatement, we compare what portion of intentional and unintentional misstatements are not classified as non-misstatements by each classifier.

Because in a multi-class task, a performance measure such as the cost curve, which directly incorporates misclassification costs, is not available (Drummond & Holte, 2000), following Zhou and Liu (2006), we create three types of cost matrices and calculate the total misclassification costs as the fourth measure of our performance evaluation. Given that actual misclassification costs are

[2] Here we do not compare our tertiary-class models with binary-class models that can be built by aggregating the two classes of misstatements (Irregularities and Errors) into one class. This comparison would not be compatible with the purpose of this study. We argue that the past studies that employ two-class models are insufficient since binary-class detection models can cause incorrect inferences if the target variable contains both intentional and unintentional misstatements or may be ineffective in detecting unintentional misstatements if only intentional instances are used to build detection models.

Table 8
Three cost matrices.

		Predicted								
		Cost matrix 1			Cost matrix 2			Cost matrix 3		
		I	U	N	I	U	N	I	U	N
Actual	Intentional (I)	0	10	50	0	10	70	0	20	100
	Unintentional (U)	10	0	30	10	0	20	10	0	50
	No misstated (N)	5	5	0	10	10	0	10	10	0

Table 9
Classification result.

Classifier	Accuracy	G-mean	% Misstatments detected	Cost 1	Cost 2	Cost 3
MLogit	0.884	0.605	75.2%	100.00	100.00	100.00
SVM-Lin	0.877	0.580	75.4%	100.22	98.10	99.83
BayesNet	0.821	0.550	68.7%	130.02	136.90	129.17

Table 10
Classification result using MetaCost.

Classifier	Accuracy	G-mean	% Misstatments detected	Cost 1	Cost 2	Cost 3
MLogit	0.869	0.698	92%	69.3	85.9	66.1
SVM-Lin	0.854	0.656	90%	73.8	89.9	70.8
BayesNet	0.825	0.584	76%	113.7	94.8	120.5

Table 11
Feature values by class.

Class	Sales	Assets	Leverage		Accruals	
	median	median	Median	Std. Dev.	Median	Std. Dev.
Intentional	247.86	352.18	0.31	0.68	0.06	0.33
Unintentional	206.66	209.95	0.21	**1.99**	0.04	**3.53**
No misstated	261.36	257.46	0.23	0.46	0.00	0.38

different for different users, we select three cost sets within the cost range of previous studies on two-class classification such as Beneish (1999) (see Table 8).

The classification result is presented in Table 9. For ease of comparison, costs are normalized by costs produced by MLogit. Multinomial logistic regression and support vector machine show comparable results. We also undertake a cost-sensitive learning using the cost matrices in Table 8. The result is shown in Table 10, where the first three columns (accuracy, G-mean and % Misstatements detected) show the maximum value under the three different cost scenarios.

With the MetaCost procedure, we document that G-mean measures and misstatement detection rates improved significantly. As shown earlier, multinomial logistic regression and support vector machine show fairly comparable results when compared under each cost scenario.

The G-mean values are not very satisfactory because it is difficult to classify unintentional-misstatement instances correctly. As shown in Table 11, the unintentional-misstatement class consists of smaller firms in terms of sales and total assets. They are less leveraged (less debt and thus a safer investment). Yet, in terms of accruals, a measure of earnings quality, the unintentional class shows higher accrual values than non-misstated firms but lower values than intentional-misstatement firms. However, what actually causes low G-mean score is variability. The unintentional group shows a much higher standard deviation for many features (not all variables reported here), making classification much more challenging. This suggests that it is necessary to break down the unintentional class further into a number of separate classes with similar characteristics, such as misstatement causes, to improve the classification results.

Another possible explanation for low G-mean score is that our target variable is not a perfect measure of managerial intent. As Hennes et al. (2008) noted in their study, management intent is impossible to observe in actuality. In our study, we assume that the target variable was correctly labeled but, due to the inability to measure human intention perfectly, some instances may have been mislabeled, causing less accurate detection rates.

Moreover, it is noteworthy that a multiclass classification problem is intrinsically more complex than a binary problem, since the generated classifier must be able to separate the data into a higher number of categories, which increases the chances of classification errors. As a result, its complexity increases for more classes (Liu, Ranka, & Kahveci, 2008; Lorena, De Carvalho, & Gama, 2008). Especially when the data are of high dimensionality and the sample size is small, the classification accuracy degrades very rapidly as the number of class increases (Li, Zhang, & Ogihara, 2004). Our tertiary models are theoretically better construct than the binary models employed by the past studies but G-means (accuracy per class) may be lower compared to the binary models.

5. Conclusion and future research

Financial fraud has substantial economic consequences for any economy. Management that intentionally misleads users of their financial statements will invoke considerable financial and social costs. While researchers have extensively analyzed the causes, motivations, and consequences of financial misstatements and earnings manipulation, we are unaware of any research that predicts accounting fraud by managerial intent. Hennes et al. (2008) suggests that it is important to distinguish intentional from unintentional misstatements. Unlike unintentional financial misstatements,

intentional financial misstatements are likely to cause severe problems and in turn shake the confidence in our financial markets.

Using the post-event analysis in Hennes et al. (2008), we develop three-class financial misstatement detection models. The models are developed to detect financial misstatements and classify misstatements according to fraud intention. To the best of our knowledge, the present study is the first multi-class predictive study that attempts to detect and classify financial misstatements according to fraud intention. We also apply multi-class cost-sensitive learning using MetaCost to deal with class imbalances and asymmetric misclassification costs.

Variables related to accruals quality, such as changes in inventory along with industry-level and organizational context-based measures, have shown discriminatory power. The firm-efficiency measure and market variables such as the short interest ratio are also found to be useful to detect misstatements and deliberate fraud.

We acknowledge that management intention is not observable; thus, our target variable may not be a perfect measure of managerial intent. Also, building multi-class detection models is more difficult than building binary class models because of higher complexity in the definition of the decision boundaries associated with larger number of classes. Nonetheless, understanding and detecting fraud intention is a crucial step toward preventing misstatements effectively.

To improve our models further, a breakdown of unintentional-misstatement firms may be necessary. Also, future studies can incorporate audit-related variables such as audit fees and internal control variables to improve detectability even further.

Acknowledgments

This work was supported by the BK21 Plus Program (Center for Sustainable and Innovative Industrial Systems, Dept. of Industrial Engineering, Seoul National University(SNU)) funded by the Ministry of Education, Korea (No. 21A20130012638), the National Research Foundation(NRF) grant funded by the Korea government(MSIP) (No. 2011-0030814), and the Institute for Industrial Systems Innovation of SNU.

Bok Baik acknowledges financial support from Institute of Management Research, Seoul National University.

References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: a meta-learning framework for detecting financial fraud. *Mis Quarterly, 36*(4), 1293–1327.

Badertscher, B. A. (2011). Overvaluation and the choice of alternative earnings management mechanisms. *The Accounting Review, 86*(5), 1491–1518.

Baik, B., Chae, J., Choi, S., & Farber, D. B. (2013). Changes in operational efficiency and firm performance: A frontier analysis approach. *Contemporary Accounting Research, 30*(3), 996–1026.

Beasley, M. S. (1996). An empirical analysis of the relation between the board of director composition and financial statement fraud. *Accounting Review*, 443–465.

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal, 55*(5), 24–36.

Beneish, M. D. (1999). Incentives and penalties related to earnings overstatements that violate gaap. *The Accounting Review, 74*(4), 425–457.

Bermejo, P., Gámez, J. A., & Puerta, J. M. (2011). Improving the performance of naive bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. *Expert Systems with Applications, 38*(3), 2072–2080.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, 235–249.

Bourke, C., Deng, K., Scott, S. D., Schapire, R. E., & Vinodchandran, N. (2008). On reoptimizing multi-class classifiers. *Machine Learning, 71*(2–3), 219–242.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science, 56*(7), 1146–1160.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications, 41*(10), 4915–4928.

Dechow, P., Ge, W., & Schrand, C. (2010). Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics, 50*(2), 344–401.

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary accounting research, 28*(1), 17–82.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *Accounting review*, 193–225.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the sec. *Contemporary accounting research, 13*(1), 1–36.

DeFond, M. L., & Jiambalvo, J. (1994). Debt covenant violation and manipulation of accruals. *Journal of accounting and economics, 17*(1), 145–176.

Demerjian, P., Lev, B., & McVay, S. (2012). Quantifying managerial ability: A new measure and validity tests. *Management Science, 58*(7), 1229–1248.

Demerjian, P. R., Lev, B., Lewis, M. F., & McVay, S. E. (2012). Managerial ability and earnings quality. *The Accounting Review, 88*(2), 463–498.

Dichev, I. D., Graham, J. R., Harvey, C. R., & Rajgopal, S. (2013). Earnings quality: Evidence from the field. *Journal of Accounting and Economics, 56*(2), 1–33.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth acm sigkdd international conference on knowledge discovery and data mining* (pp. 155–164). ACM.

Drummond, C., & Holte, R. C. (2000). Explicitly representing expected cost: An alternative to roc representation. In *Proceedings of the sixth acm sigkdd international conference on knowledge discovery and data mining* (pp. 198–207). ACM.

Erickson, M., Hanlon, M., & Maydew, E. L. (2006). Is there a link between executive equity incentives and accounting fraud? *Journal of Accounting Research*, 113–143.

Ettredge, M., Scholz, S., Smith, K. R., & Sun, L. (2010). How do restatements begin? evidence of earnings management preceding restated financial reports. *Journal of Business Finance & Accounting, 37*(3–4), 332–355.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery, 1*(3), 291–316.

GAO (2002). *Financial statement restatements: Trends, market impacts, regulatory responses, and remaining challenges*. US General Accounting Office.

Gillett, P. R., & Uddin, N. (2005). Cfo intentions of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory, 24*(1), 55–75.

GMI Ratings (2013). The gmi ratings agr model: Measuring accounting and governance risk in public corporations. http://www3.gmiratings.com/wp-content/uploads/2013/11/GMIRatings_AGR3.0Whitepaper_102013.pdf.

Green, B. P., & Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing, 16*(1), 14.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques: Concepts and techniques*. Elsevier.

Hastie, T., Tibshirani, R., et al. (1998). Classification by pairwise coupling. *The annals of statistics, 26*(2), 451–471.

Hayes, L. (2014). Identifying unintentional error in restatement disclosures. Available at SSRN 2269086,.

Hennes, K. M., Leone, A. J., & Miller, B. P. (2008). The importance of distinguishing errors from irregularities in restatement research: The case of restatements and ceo/cfo turnover. *The Accounting Review, 83*(6), 1487–1519.

Hennes, K. M., Leone, A. J., & Miller, B. P. (2013). Determinants and market consequences of auditor dismissals after accounting restatements. *The Accounting Review, 89*(3), 1051–1082.

Hoitash, R., Hoitash, U., & Johnstone, K. M. (2012). Internal control material weaknesses and cfo compensation. *Contemporary Accounting Research, 29*(3), 768–803.

Huang, S.-Y., Tsaih, R.-H., & Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert Systems with Applications, 41*(9), 4360–4372.

Jones, K. L., Krishnan, G. V., & Melendrez, K. D. (2008). Do models of discretionary accruals detect actual cases of fraudulent and restated earnings? an empirical analysis. *Contemporary Accounting Research, 25*(2), 499–531.

Kim, J., Choi, K., Kim, G., & Suh, Y. (2012). Classification cost: An empirical comparison among traditional classifier, cost-sensitive classifier, and metacost. *Expert Systems with Applications, 39*(4), 4013–4019.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*(4), 995–1003.

Kotsiantis, S., Koumanakos, E., Tzelepis, D., & Tampakas, V. (2006). Forecasting fraudulent financial statements using data mining. *International Journal of Computational Intelligence, 3*(2), 104–110.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning, 59*(1–2), 161–205.

Leong, C. K. (2015). Credit risk scoring with bayesian network models. *Computational Economics*, 1–24.

Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics, 20*(15), 2429–2437.

Lin, C.-C., Chiu, A.-A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and expertsâ?? judgments. *Knowledge-Based Systems, 89*, 459–470.

Ling, C. X., & Sheng, V. S. (2010). Cost-sensitive learning. In *Encyclopedia of machine learning* (pp. 231–235). Springer.

Liu, J., Ranka, S., & Kahveci, T. (2008). Classification and feature selection algorithms for multi-class cgh data. *Bioinformatics, 24*(13), i86–i95.

Liu, X.-Y., & Zhou, Z.-H. (2006). The influence of class imbalance on cost-sensitive learning: An empirical study. In *Data mining, 2006. icdm'06. sixth international conference on* (pp. 970–974). IEEE.

Lorena, A. C., De Carvalho, A. C., & Gama, J. M. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review, 30*(1-4), 19–37.

McVay, S. E. (2006). Earnings management using classification shifting: An examination of core earnings and special items. *The Accounting Review, 81*(3), 501–531.

Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569.

Pai, P.-F., Hsu, M.-F., & Wang, M.-C. (2011). A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems, 24*(2), 314–321.

Palmrose, Z.-V., Richardson, V. J., & Scholz, S. (2004). Determinants of market reactions to restatement announcements. *Journal of accounting and economics, 37*(1), 59–89.

Palmrose, Z.-V., & Scholz, S. (2004). The circumstances and legal consequences of non-gaap reporting: Evidence from restatements. *Contemporary Accounting Research, 21*(1), 139–180.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19–50.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers, 10*(3), 61–74.

Plumlee, M., & Yohn, T. L. (2010). An analysis of the underlying causes attributed to restatements. *Accounting Horizons, 24*(1), 41–64.

Ragothaman, S., Carpenter, J., & Buttars, T. (1995). Using rule induction for knowledge acquisition: An expert systems approach to evaluating material errors and irregularities. *Expert systems with Applications, 9*(4), 483–490.

Richardson, S. A., Sloan, R. G., Soliman, M. T., & Tuna, I. (2005). Accrual reliability, earnings persistence and stock prices. *Journal of accounting and economics, 39*(3), 437–485.

Roychowdhury, S. (2006). Earnings management through real activities manipulation. *Journal of accounting and economics, 42*(3), 335–370.

Schrand, C. M., & Zechman, S. L. (2012). Executive overconfidence and the slippery slope to financial misreporting. *Journal of Accounting and Economics, 53*(1), 311–329.

Sun, Y., Kamel, M. S., & Wang, Y. (2006). Boosting for learning multiple classes with imbalanced class distribution. In *Data mining, 2006. icdm'06. sixth international conference on* (pp. 592–602). IEEE.

Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems, 75*, 38–48.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security, 57*, 47–66.

Whiting, D. G., Hansen, J. V., McDonald, J. B., Albrecht, C., & Albrecht, W. S. (2012). Machine learning methods for detecting patterns of management fraud. *Computational Intelligence, 28*(4), 505–527.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xia, F., Yang, Y.-w., Zhou, L., Li, F., Cai, M., & Zeng, D. D. (2009). A closed-form reduction of multi-class cost-sensitive learning to weighted multi-class learning. *Pattern Recognition, 42*(7), 1572–1581.

Zhou, Z.-H., & Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *Knowledge and Data Engineering, IEEE Transactions on, 18*(1), 63–77.

Zhou, Z.-H., & Liu, X.-Y. (2010). On multi-class cost-sensitive learning. *Computational Intelligence, 26*(3), 232–257.