Decision Support Systems 50 (2011) 559–569

Contents lists available at ScienceDirect

Decision Support Systems

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

The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature

E.W.T. Ngai [a], Yong Hu [b,*], Y.H. Wong [a], Yijun Chen [b], Xin Sun [b]

[a] Department of Management and Marketing, The Hong Kong Polytechnic University, Kowloon, Hong Kong, PR China
[b] Institute of Business Intelligence and Knowledge Discovery, Department of E-commerce, Guangdong University of Foreign Studies, Sun Yat-Sen University, Guangzhou 510006, PR China

ARTICLE INFO

Available online 19 August 2010

Keywords:
Financial fraud
Fraud detection
Literature review
Data mining
Business intelligence

ABSTRACT

This paper presents a review of — and classification scheme for — the literature on the application of data mining techniques for the detection of financial fraud. Although financial fraud detection (FFD) is an emerging topic of great importance, a comprehensive literature review of the subject has yet to be carried out. This paper thus represents the first systematic, identifiable and comprehensive academic literature review of the data mining techniques that have been applied to FFD. 49 journal articles on the subject published between 1997 and 2008 was analyzed and classified into four categories of financial fraud (bank fraud, insurance fraud, securities and commodities fraud, and other related financial fraud) and six classes of data mining techniques (classification, regression, clustering, prediction, outlier detection, and visualization). The findings of this review clearly show that data mining techniques have been applied most extensively to the detection of insurance fraud, although corporate fraud and credit card fraud have also attracted a great deal of attention in recent years. In contrast, we find a distinct lack of research on mortgage fraud, money laundering, and securities and commodities fraud. The main data mining techniques used for FFD are logistic models, neural networks, the Bayesian belief network, and decision trees, all of which provide primary solutions to the problems inherent in the detection and classification of fraudulent data. This paper also addresses the gaps between FFD and the needs of the industry to encourage additional research on neglected topics, and concludes with several suggestions for further FFD research.

1. Introduction

In recent years, financial fraud, including credit card fraud, corporate fraud and money laundering, has attracted a great deal of concern and attention. The Oxford English Dictionary [55], p. 562] defines fraud as "wrongful or criminal deception intended to result in financial or personal gain." Phua et al. [58] describe fraud as leading to the abuse of a profit organization's system without necessarily leading to direct legal consequences. Although there is no universally accepted definition of financial fraud, Wang et al. [78], p. 1120] define it as "a deliberate act that is contrary to law, rule, or policy with intent to obtain unauthorized financial benefit."

Economically, financial fraud is becoming an increasingly serious problem. A striking case in point is the Ponzi scheme perpetuated by former NASDAQ chairman Bernard Madoff, which has led to the loss of approximately US$50 billion worldwide [34]. Another example is that of Joseph Hirko, former co-chief executive officer of Enron Broadband Services (EBS), who has avowed to forfeit approximately US $8.7 million in restitution to Enron victims through the U.S. Securities and Exchange Commission's Enron Fair Fund after pleading guilty to wire fraud [34]. According to a 2007 BBC news report [8], fraudulent insurance claims cost UK insurers a total of 1.6 billion pounds a year. The overall losses caused by financial fraud are incalculable.

Financial fraud detection (FFD) is vital for the prevention of the often devastating consequences of financial fraud. FFD involves distinguishing fraudulent financial data from authentic data, thereby disclosing fraudulent behavior or activities and enabling decision makers to develop appropriate strategies to decrease the impact of fraud.

Data mining plays an important role in FFD, as it is often applied to extract and uncover the hidden truths behind very large quantities of data. Bose and Mahapatra [14] define data mining as a process of identifying interesting patterns in databases that can then be used in decision making. Turban et al. [73] define data mining as a process that uses statistical, mathematical, artificial intelligence, and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from a large database. Frawley et al. [35] state that the objective of data mining is to obtain useful, non-explicit information from data stored in large repositories. Kou et al. [47] highlight that an important advantage of data mining is that it can be used to develop a new class of models to identify new attacks before

* Corresponding author. Tel.: +86 20 39328876.
E-mail addresses: mswtngai@polyu.edu.hk (E.W.T. Ngai), huyong2@mail2.sysu.edu.cn (Y. Hu).

0167-9236/$ – see front matter.
doi:10.1016/j.dss.2010.08.006

they can be detected by human experts. Phua et al. [58] point out that fraud detection has become one of the best established applications of data mining in both industry and government. Various data mining techniques have been applied in FFD, such as neural networks [18,27,31,38,45,75], logistic regression models [10,54,65,85], the naïve Bayes method [11,77], and decision trees [45,46], among others.

Over the past few years, a number of review articles have appeared in conference or journal publications. Bolton and Hand [13], for example, have reviewed statistical methods of detecting fraud, including credit card fraud, money laundering, telecommunications fraud, etc. Zhang and Zhou [88] have surveyed financial applications of data mining including stock market and bankruptcy predictions and fraud detection. Phua et al. [58] present a survey of data mining-based fraud detection research, including credit transaction fraud, telecoms subscription fraud, automobile insurance fraud and the like. Others have reviewed insurance fraud [24] and financial statement fraud [86]. However, the survey presented herein is an up-to-date, comprehensive and state-of-the-art review of data mining applications in FFD.

This paper has three objectives. The first is to develop a framework for classifying the applications of data mining to FFD. The second is to provide a systematic and comprehensive review of existing research articles on the applications of data mining to FFD. The third is to use the review and framework to generate a roadmap for researchers and practitioners seeking to better comprehend this field.

The remainder of this article is structured as follows. Section 2 presents the methodological framework for research. Section 3 provides our classification framework for the application of data mining in FFD. Section 4 analyzes FFD research according to this classification framework. Section 5 concludes our research and suggests further research directions.

2. Methodological framework for research

The methodological framework for this research can be divided into three essential phases: research definition, research methodology, and research analysis, as depicted in Fig. 1.

In phase 1, we determine the research area, the expected research goal, and the research scope. The research area is academic research on FFD that applies data mining techniques. The research goal is to create a classification framework for the data mining techniques applied to FFD and to suggest directions for future research. The research scope is the literature on the applications of data mining to FFD published between 1997 and 2008, which is summarized to aid the further creation and accumulation of knowledge in this area. As the research on this topic is relatively recent, the scope of this investigation is limited to the time frame of 1997 to 2008, but this 12-year period is deemed to be representative of the application of data mining to FFD.

In phase 2, we define the criteria for searching for and selecting articles, and create a framework to classify the selected articles. Nine online academic databases were searched to provide a comprehensive listing of journal articles, as the nature of FFD and data mining research makes it difficult to confine the search to specific disciplines. These databases cover most academic journals in English available in full text versions.

- ABI/INFORM Database
- Academic Search Premier
- ACM
- Business Source Premier
- Emerald Full text
- IEEE Transactions
- Science Direct
- Springer-Link Journals
- World Scientific Net

This literature search was based on the descriptors “financial fraud,” “data mining” and “business intelligence.” We used Boolean expressions to apply these terms to a search of online databases, which originally produced approximately 1200 articles. The review and classification process was carefully and independently verified by the co-authors, and only articles that were related to data mining and FFD were included. Each article was carefully examined to ensure that it met the three selection criteria. First, the articles must have been published in academic journals for which the full text versions are available. Conference articles, master or doctoral dissertations, textbooks, and unpublished working papers were excluded, largely for reasons of availability. Second, the articles had to have been published

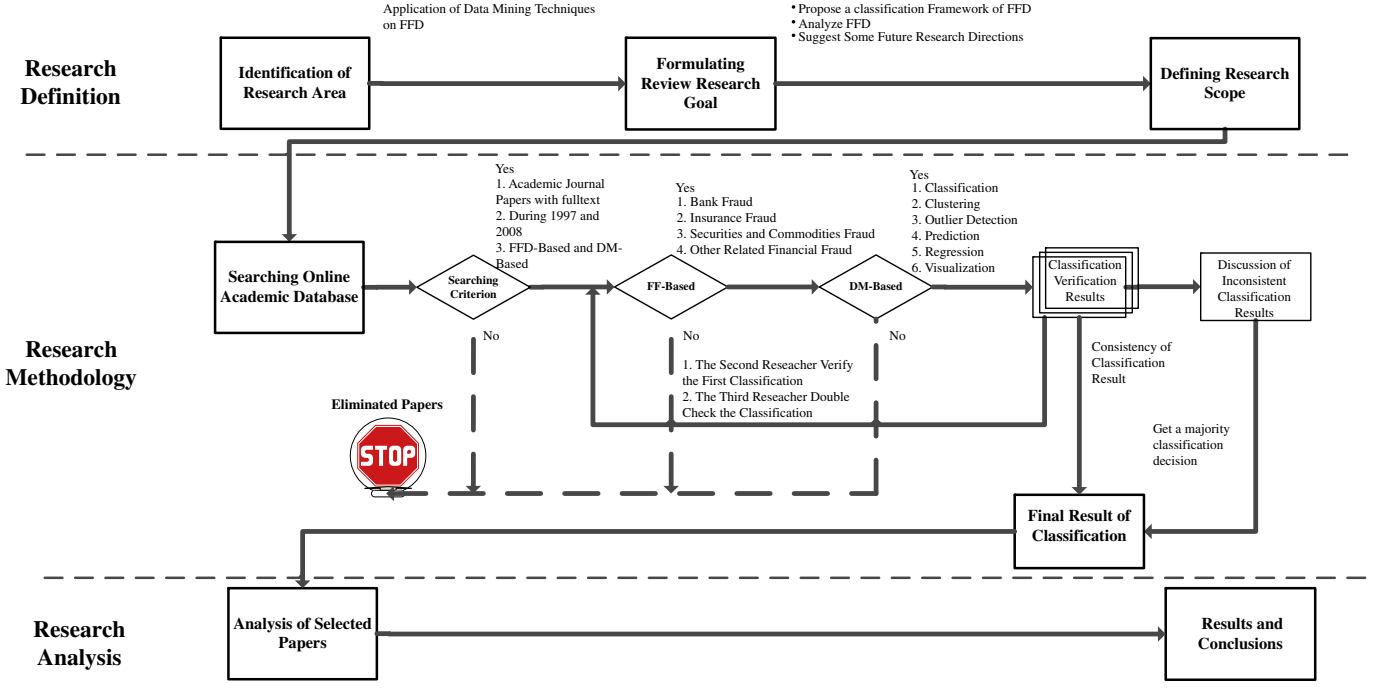

Fig. 1. Methodological framework for research.

between 1997 and 2008. Third, the articles had to present data mining techniques and discuss their application to financial fraud.

Forty-nine articles were selected for classification. Each was classified according to the following steps [53].

- Classify the articles selected by one of the co-authors.
- Verify the classification with another co-author and double check with another independent co-author.
- Approve the categories assigned to the article if the classification results are consistent, or hold a discussion among the researchers to reach a consensus otherwise.

In the last phase, we analyzed the selected articles to draw some conclusions and identify some future research directions. The details of the analysis are presented in Section 4.

3. Classification framework on data mining and financial fraud detection

In this section, we propose a graphical conceptual classification framework for the available literature on the applications of data mining techniques to FFD. The classification framework, which is shown in Fig. 2, is based on a literature review of existing knowledge on the nature of data mining research [3,52], fraud detection research [13,24,58,86,88], and the financial crime framework of the U.S. Federal Bureau of Investigation [33] which is summarized and presented in Table 1.

Our proposed classification framework for financial fraud is based on the financial crime framework of the U.S. Federal Bureau of Investigation [33], because it is one of the best established frameworks for FFD. The classification of financial fraud comprises two levels, as shown in Table 1. The higher level comprises *financial fraud based (FF-based)* categories, which include bank fraud, insurance fraud, securities and commodities fraud, and other related financial fraud, whereas the lower level comprises *fraudulent activities*, including mortgage fraud, asset forfeiture/money laundering, healthcare fraud, insurance fraud, securities and commodities fraud, corporate fraud, and mass marketing fraud.

Fig. 2 consists of two layers, the first comprising the aforementioned financial fraud based categories and the second comprising the six *data mining application classes* of classification, clustering, prediction, outlier detection, regression, and visualization [13,24,32,48,58,61,74,76], supported by a set of algorithmic approaches to extract the relevant relationships in the data [73]. We provide a brief description of our conceptual framework with references, and of the six data mining application classes (classification, clustering, outlier detection, prediction, regression and visualization), each component of which is discussed in more detail in the following sections.

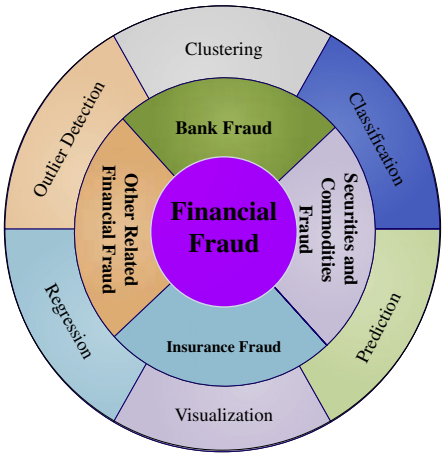

Fig. 2. Conceptual framework for classifying the applications of data mining to FFD.

Table 1
Classification for financial fraud based on FBI [33].

Financial fraud based categories	Fraudulent activities
Bank fraud	Mortgage fraud, Asset forfeiture/money laundering
Insurance fraud	Healthcare fraud, Insurance fraud
Securities and commodities fraud	Securities and commodities fraud
Other related financial fraud	Corporate fraud, Mass marketing fraud

3.1. Classification for financial fraud

As previously mentioned, in this study, financial fraud is classified into four broad categories. They are:

Bank fraud. According to Connell University Law School (CULS) [22], bank fraud is defined as "whoever knowingly executes, or attempts to execute, a scheme or artifice (1) to defraud a financial institution; or (2) to obtain any of the moneys, funds, credits, assets, securities, or other property owned by, or under the custody or control of, a financial institution, by means of false or fraudulent pretenses, representations, or promises."
In this study, bank fraud includes credit card fraud, money laundering, and mortgage fraud, where mortgage fraud is defined as "material misstatement, misrepresentation, or omission relating to the property or potential mortgage relied on by an underwriter or lender to fund, purchase or insure a loan" [33] and credit card fraud is defined as the unauthorized usage of a card, unusual transaction behavior, or transactions on an inactive card [70]. According to the FBI [33], money laundering is the process by which criminals conceal or disguise the proceeds of their crimes or convert those proceeds into goods and services. It allows criminals to inject their illegal money into the stream of commerce, thus corrupting financial institutions and the money supply and giving criminals unwarranted economic power. Gao and Ye [36] similarly define money laundering as the process by which criminals "wash dirty money" to disguise its illicit origin and make it appear legitimate and "clean."
Insurance fraud. Insurance fraud can occur at many points in the insurance process (e.g., application, eligibility, rating, billing, and claims), and can be committed by consumers, agents and brokers, insurance company employees, healthcare providers, and others [21,44]. In this study, insurance fraud includes crop, healthcare, and automobile insurance fraud. FBI [33] states that healthcare fraud is carried out by many segments of the healthcare system using various methods, with some of the most prevalent types of fraud including "Billing for Services not Rendered, Upcoding of Services, Upcoding of Items, Duplicate Claims, Unbundling, Excessive Services, Medically Unnecessary Services and Kickbacks" [33]. Crop insurance fraud is committed by purchasers of crop insurance who fake or overstate either the loss of their crops due to natural disasters or the loss of revenue due to declines in the price of agricultural commodities. Automobile insurance fraud comprises a set of fraudulent activities that include staged accidents, superfluous repairs, and faked personal injuries.
Securities and commodities fraud. The FBI [33] provides brief descriptions of some of the most prevalent securities and commodities frauds encountered today, for example, "Market Manipulation, High Yield Investment Fraud, The Ponzi Scheme,

The Pyramid Scheme, Prime Bank Scheme, Advance Fee Fraud, Hedge Fund Fraud, Commodities Fraud, Foreign Exchange Fraud, Broker Embezzlement and Late-Day Trading." According to another definition by CULS [22], securities frauds include theft from manipulation of the market, theft from securities accounts, and wire fraud.

Other related financial fraud. Our final category is made up of types of financial fraud other than those in the aforementioned categories, such as corporate fraud and mass marketing fraud. Again, according to FBI [33], "corporate fraud investigations involve the following activities: (1) falsification of financial information, (2) self-dealing by corporate insiders, and (3) obstruction of justice designed to conceal any of the above-noted types of criminal conduct." The Bureau further states that "mass marketing fraud is a general term for types of fraud that exploit mass-communication media, such as telemarketing, mass mailings, and the Internet."

3.2. Classification of data mining applications and techniques

Each of the six data mining application classes is supported by a set of algorithmic approaches to extract the relevant relationships in the data [73]. These approaches differ in the classes of problems that they are able to solve (see [40]). The classes are as follows.

Classification. Classification builds up and utilizes a model to predict the categorical labels of unknown objects to distinguish between objects of different classes. These categorical labels are predefined, discrete and unordered [39,71]. Zhang and Zhou [88] state that classification and prediction is the process of identifying a set of common features and models that describe and distinguish data classes or concepts. Common classification techniques include neural networks, the naïve Bayes technique, decision trees and support vector machines. Such classification tasks are used in the detection of credit card, healthcare and automobile insurance, and corporate fraud, among other types of fraud, and classification is one of the most common learning models in the application of data mining in FFD.

Clustering. Clustering is used to divide objects into conceptually meaningful groups (clusters), with the objects in a group being similar to one another but very dissimilar to the objects in other groups. Clustering is also known as data segmentation or partitioning and is regarded as a variant of unsupervised classification [39,71]. According to Yue et al. [86], p. 5520], "clustering analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups." Further, Zhang and Zhou [88] argue that each cluster is a collection of data objects which are similar to one another within the same cluster but dissimilar to those in other clusters. The most common clustering techniques are the K-nearest neighbor, the Naïve Bayes technique and self-organizing map techniques.

Prediction. Prediction estimates numeric and ordered future values based on the patterns of a data set [3,12]. Han and Kamber [39] note that, for prediction, the attribute for which the values are being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered). This attribute can be referred to simply as the predicted attribute. Neural networks and logistic model prediction are the most commonly used prediction techniques.

Outlier detection. Outlier detection is employed to measure the "distance" between data objects to detect those objects that are grossly different from or inconsistent with the remaining data set [39]: "Data that appear to have different characteristics than the rest of the population are called outliers" [2], p. 521]. Yamanishi et al. [82] point out that the problem of outlier/anomaly detection is one of the most fundamental issues in data mining. A commonly used technique in outlier detection is the discounting learning algorithm.

Regression. Regression is a statistical methodology used to reveal the relationship between one or more independent variables and a dependent variable (that is continuous-valued) [39]. Many empirical studies have used logistic regression as a benchmark [1,28,62,76,79]. The regression technique is typically undertaken using such mathematical methods as logistic regression and linear regression, and it is used in the detection of credit card, crop and automobile insurance, and corporate fraud.

Visualization. Visualization refers to the easily understandable presentation of data and to methodology that converts complicated data characteristics into clear patterns to allow users to view the complex patterns or relationships uncovered in the data mining process [63,73]. Eick and Fyock [29] report that researchers at Bell and AT&T Laboratories have exploited the pattern detection capabilities of the human visual system by building a suite of tools and applications that flexibly encode data using color, position, size and other visual characteristics. Visualization is best used to deliver complex patterns through the clear presentation of data or functions.

4. Analysis of FFD research based on the proposed classification framework

This paper provides a state-of-the-art review of the applications of data mining to FFD. Fig. 3, which is based on Fig. 2, dissects and organizes this review of the literature. For the classification of financial fraud, we divide the articles among the categories of bank fraud, insurance fraud, securities and commodities fraud, and other related financial fraud. In the second level of the classification, we make a further categorization based on fraudulent activities (e.g., asset forfeiture/money laundering). For the data mining classification, we first identify six data mining application classes, and then in the second level of classification make a further categorization using a set of algorithmic approaches (e.g., neural networks).

The distribution of the 49 articles classified into the proposed classification framework is given in Table 2. Table 2 lists the applications of data mining to FFD by the FF-based categories and fraudulent activities, and identifies the data mining application classes and techniques used with reference to the problems addressed. Some of the selected applications in the review address more than one FDD problem, and thus we categorized these applications by the dominant problem addressed.

A complete list of the 49 selected articles is presented in Tables 3–5. The first column of Tables 3–5 present the important literature studied in our research. The second column gives a brief description of the articles and their main objectives. The following subsections present further analysis of data mining techniques in FFD.

4.1. Distribution of articles by data mining application classes

The classification of the 49 articles by data mining application classes is shown in Table 6.

Judging by the numbers of published papers (see Table 6), we can clearly see that the focus of data mining applications has most often

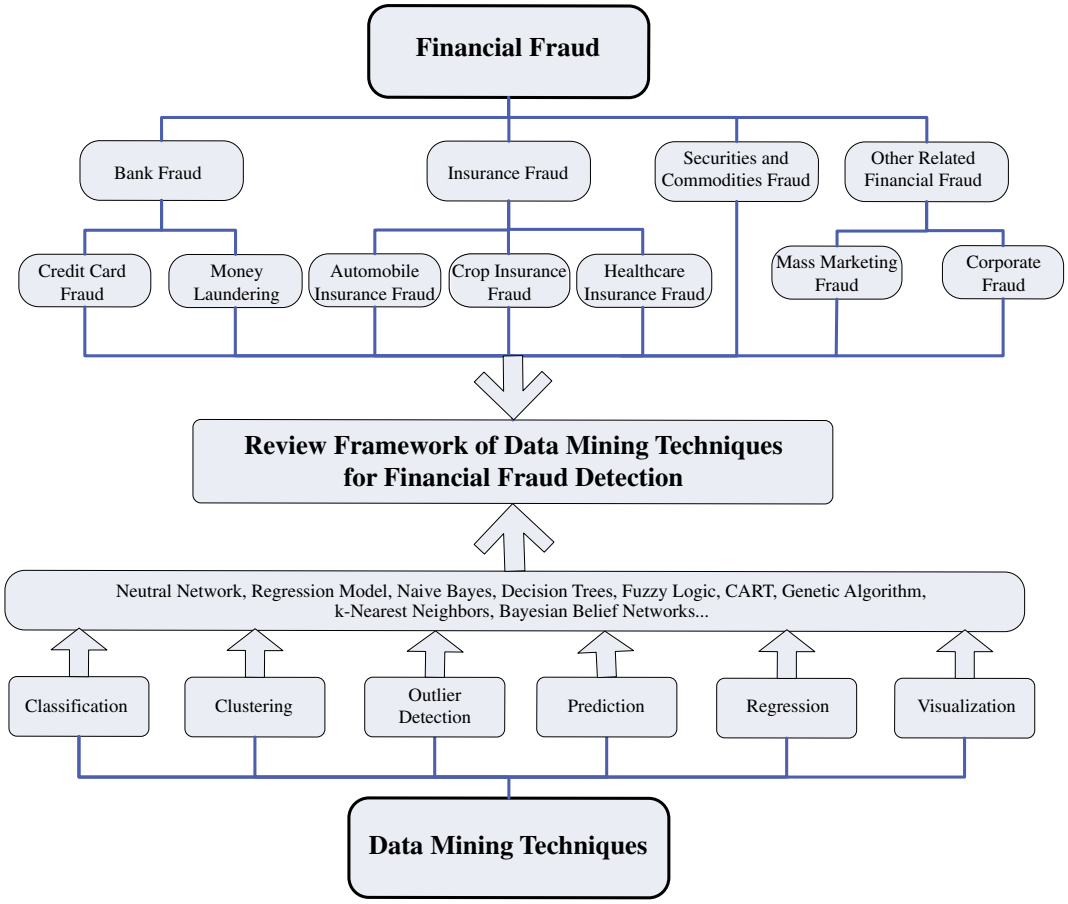

Fig. 3. Framework for dissection and organization of the review of articles.

been on automobile insurance fraud and corporate fraud (17 or 34.7% each), followed by credit card fraud (7 or 14.3%). Overall, insurance fraud is the most prominent area for the application of data mining techniques in FFD (49%). It is worth noting that there is no published article related to mortgage fraud, securities and commodities fraud and mass marketing fraud within our selection criteria; thus they are not listed in the above-mentioned table.

It can be clear that classification is the most frequently used data mining application class, accounting for 61.2% of the total (30 of the 49 articles), and that outlier detection and visualization are the least common, accounting for only 2.0% each (1 out of 49 each). Given that outlier detection is a significant method of fraud detection, which has characteristics that confer comparative advantages over other techniques, more attention should be paid to it in future research.

4.2. Distribution of articles by data mining techniques

To determine the main algorithms used for FFD, we present a simple analysis of FFD and the data mining techniques identified in the articles in Table 7. Twenty-six data mining techniques have been applied to the detection of financial fraud. Mortgage fraud, securities and commodities fraud, and mass marketing fraud are not listed in the table because the techniques identified in our research have not been applied to these problems. The most frequently used techniques are logistic models, neural networks, the Bayesian belief network, and decision trees, all of which fall into the "classification" category. Of these techniques, logistic models are the most popular, being used in 21.3% (16 of 75) of the studies reviewed, followed by neural networks, used in 13.3% (10 of 75), and then the Bayesian belief network and decision trees, both used in 6.7% (5 of 75) of studies. These four techniques are discussed in more detail in the following paragraphs.

Logistic model. Logistic model is a generalized linear model that is used for binomial regression in which the predictor variables can be either numerical or categorical [65,84]. It is principally used to solve problems caused by automobile insurance and corporate fraud.

Neural networks. The neural network is a technique that imitates the functionality of the human brain using a set of interconnected vertices [37,84]. It is widely applied in classification and clustering, and its advantages are as follows. First, it is adaptive; second, it can generate robust models; and third, the classification process can be modified if new training weights are set. Neural networks are chiefly applied to credit card, automobile insurance and corporate fraud.

Bayesian belief network. The Bayesian belief network (BBN) represents a set of random variables and their conditional independencies using a directed acyclic graph (DAG), in which nodes represent random variables and missing edges encode conditional independencies between the variables [45,57]. The Bayesian belief network is often adopted in credit card, automobile insurance, and corporate fraud detection.

Decision trees. Decision trees are predictive decision support tools that create mapping from observations to possible consequences [39,49]. These trees can be planted via machine-learning-based

Table 2
Research on data mining techniques in FFD.

FF-based categories	Fraudulent activities	Data mining application class	Data mining techniques	References
Bank fraud	Credit card fraud	Classification	Ada boost algorithm, decision trees, CART, RIPPER, Bayesian Belief Network,	[19]
			Neural networks, discriminant analysis	[27]
			K-nearest neighbor, logistic model, discriminant analysis, Naïve Bayes, neural networks, decision trees	[84]
			Support vector machine, evolutionary algorithms	[20]
		Clustering	Hidden Markov Model	[67]
			Self-organizing map	[60,87]
	Money laundering	Classification	Network analysis	[36]
Insurance fraud	Crop insurance fraud	Regression	Yield-switching model	[6]
			Logistic model, probit model	[43]
	Healthcare insurance fraud	Classification	Association rule	[83]
			Polymorphous (M-of-N) logic	[51]
			Self-organizing map	[41]
		Visualization	Visualization	[64]
		Outlier detection	Discounting learning algorithm	[82]
	Automobile insurance fraud	Classification	Logistic model	[17]
			Neural networks	[75]
			Principal component analysis of RIDIT(PRIDIT)	[16]
			Logistic model	[74]
			Logistic model, decision trees, neural networks, support vector machine, K-nearest neighbor, Naïve Bayes, Bayesian belief network	[76]
			Fuzzy logic	[56]
			Logistic model	[4,5]
			Logistic model, Bayesian belief network	[11]
			Self-organizing map	[15]
			Naïve Bayes	[77]
		Prediction	Evolutionary algorithms	[68]
			Logistic model	[72]
		Regression	Probit model	[59]
			Logistic model	[23,80]
			Probit model	[9]
Other related financial fraud	Corporate fraud	Classification	Neural networks, decision trees, Bayesian belief network	[45]
			Multicriteria decision aid (MCDA), UTilite's Additives DIScriminantes (UTADIS)	[66]
			Evolutionary algorithms	[81]
			Fuzzy logic	[25,26]
			Neural networks	[31,38]
			Neural networks, logistic model	[50]
			Logistic model	[10]
			CART	[7]
			Decision trees, neural networks, Bayesian belief network, K-nearest neighbor, RIPPER, support vector machine, stacking variant methodology	[46]
		Clustering	Naïve Bayes	[42]
		Prediction	Neural networks	[18]
		Regression	Logistic model	[85]
			Logistic model	[65]
			Logistic model	[30]

Table 3
Bank fraud.

Reference	Main objectives
[19]	To use Ada Boost, C4.5, CART, Ripper, Bayes and ID3 methods to determine whether combining multiple learned fraud detectors under a "cost model" could reduce losses from fraud
[20]	To use a binary support vector system (BSVS) based on the support vectors in support vector machines (SVM) and the genetic algorithm (GA) to solve problems of credit card fraud that had not been well identified
[27]	To present an on-line system for fraud detection in credit card operations based on a neural classifier
[36]	To propose a framework for data mining-based network analysis in anti-money laundering research
[60]	To focus on real-time fraud detection and present a new model based on self-organizing maps to better understand spending patterns
[67]	To build a Hidden Markov Model for the sequence of operations in credit card transaction processing
[84]	To consider the case of customer default payments in Taiwan and compare the predictive accuracy of the probability of default among six data mining methods: K-nearest neighbor, logistic regression, discriminant analysis, Naïve Bayesian, neural networks and classification trees
[87]	To propose a self-organizing map algorithm to create a model of typical cardholder behavior and to analyze deviations in transactions, thus identifying suspicious ones

algorithms such as the ID3, CART and C4.5. Predictions are represented by leaves, and the conjunctions of features by branches. Decision trees are commonly used in credit card, automobile insurance, and corporate fraud.

4.3. Distribution of articles by year

Table 8 presents the distribution of articles by financial fraud and publication year. It can be seen from this table that research studies on corporate fraud and automobile insurance fraud are the most prominent, and we believe that this will continue to be the case.

As shown in Table 8, we identified only one application article for money laundering and none for mortgage fraud, securities and commodities fraud and mass marketing fraud. We believe that this is because of the difficulty of collecting such data for analysis and because publication of the finding may be prohibited due to the highly sensitive nature of the topic.

4.4. Distribution of articles by journal

Table 9 shows the distribution of the articles by the journal in which they appeared. The articles related to the use of data mining

Table 4
Insurance fraud.

Reference	Main objectives
[4]	To present discrete-choice models of fraudulent behavior and estimate the influence of insured and claims characteristics on the probability of fraud
[5]	To develop binary choice models for fraud detection and for the misclassification of the response variables in automobile insurance
[6]	To create predictions for a yield-switching model to identify producers whose reported yield patterns are consistent
[9]	To develop a probit model to aid insurance companies in their decision making and to ensure that they are better equipped to fight fraud
[11]	To develop an asymmetric or skewed logit model using Bayesian analysis for fraud detection in the Spanish insurance market
[15]	To apply a self-organizing feature map to classify automobile bodily injury claims fraud
[16]	To introduce the statistical and a priori classification and principal components analysis of RIDIT score (PRIDIT) methods to detect fraud in the automobile insurance industry
[17]	To build a fraud detection model based on a logit model and the EM algorithm to estimate an AAG model
[23]	To use a linear regression model to examine the optimal claims settlement strategy for a liability insurer
[41]	To propose Kohonen's self-organizing map to classify medical general practitioners who have been classified by expert consultants
[43]	To propose a score test to help in deciding whether to use a logit or a probit model in predicting insurance fraud probabilities
[51]	To build an EFD system to integrate expert knowledge with a statistical information assessment to identify cases of unusual provider behavior and to use the machine-learning method to develop new rules and improve the identification process
[56]	To develop a fuzzy-based expert system to identify and evaluate whether elements of fraud are involved in insurance claims settlements
[59]	To use a two-equation model (a bivariate probit model) for audit and fraud detection in automobile insurance
[64]	To use visualization tools to help investigators to recognize new and unusual patterns of activity, thus allowing a better understanding of the direction and use of limited health care fraud detection and investigation resources
[68]	To propose a cultural algorithm to detect fraudulent automobile insurance claims, non-fraudulent claims, false positive claims (non-fraudulent claims predicted to be fraudulent), and false negative claims
[72]	To use an econometric logistic model to investigate the role of claims auditing in the automobile insurance market
[74]	To use logistic regression to score claims and detect fraud using real-life data in the automobile insurance industry
[75]	To explore the explicative capabilities of neural network classifiers for personal injury protection in automobile insurance claims fraud detection
[76]	To use logistic regression, C4.5, neural network, least-squares support vector machine, K-nearest neighbor, Naïve Bayes and tree-augmented Naïve Bayes methods for the detection of fraud in PIP automobile insurance claims
[77]	To apply AdaBoosted Naïve Bayes scoring to insurance claims fraud
[80]	To apply a Tobit regression model to explore the potential for reducing unwarranted claims payments
[82]	To build a SmartSifter system based on the on-line unsupervised learning of a probabilistic model (using a finite mixture model) to detect outliers in an on-line process
[83]	To propose an adaptable and extendable detection model to the concept of clinical pathways to facilitate automatic and systematic construction

Table 5
Other related financial fraud.

Reference	Main objectives
[7]	To introduce classification and regression trees to identify and predict the impact of fraudulent financial statements
[10]	To develop a logistic regression model to estimate fraudulent financial reporting for an audit client
[18]	To use neural networks to predict the occurrence of corporate fraud at the management level
[25]	To provide a fuzzy sets model to assess the risk of managerial fraud
[26]	To build a rule-based fuzzy reasoning system to assess the risk of managerial fraud
[30]	To build an expert system applying the logit statistical model to enhance user engagement and increase reliance on the aid
[31]	To use neural networks to develop a model for detecting managerial fraud
[38]	To develop a neural network fraud classification model employing endogenous financial data in corporate fraud
[42]	To identify disgruntled employee systems fraud risk through Naïve Bayes text mining
[45]	To explore the effectiveness of neural networks, decision trees and Bayesian belief networks in detecting fraudulent financial statements (FFS) and to identify factors associated with FFS
[46]	To apply a hybrid decision support system using stacking variant methodology to detect FFS
[50]	To evaluate the utility of an integrated fuzzy neural network model for corporate fraud detection
[54]	To explore the logit regression model to detect corporate fraud in New Zealand
[65]	To use logistic regression to examine published data and develop a model to detect the factors associated with FFS
[66]	To explore the Multicriteria Decision Aid (MCDA) and UTilite's Additives DIScriminantes (UTADIS) for detecting FFS and identifying the factors associated with FFS
[81]	To use genetic algorithms to aid the decisions of Defense Contractor Audit Agency (DCAA) auditors when they are estimating the likelihood of contracts fraud
[85]	To employ a logistic regression model to test the effects of managerial compensation and market competition on financial fraud among listed companies in China

techniques for FFD are distributed across 29 journals that cover a wide range of fields, including information systems, auditing and finance, etc., which means that the application of such techniques for FFD has attracted considerable interest from scholars in different disciplines. The *Journal of Risk and Insurance* contained the most relevant articles (16.3%, or 8 of the 49 articles), followed by *Expert Systems with Applications* (12.2%, or 6 articles) and the *Managerial Auditing Journal* (8.2%, or 4 articles).

5. Conclusion, research implications and limitations

A critical part of any new research venture is the construction of a good classification framework and the establishment of a reference collection of relevant literature. The research area of FFD is no exception. Although the importance of data mining techniques in the detection of financial fraud has been recognized, a comprehensive classification framework or a systematic review of their application in

Table 6
Distribution of articles by data mining application classes.

FF-based categories	Fraudulent activities	Data mining application classes	Amount
Bank fraud	Credit card fraud		7 (14.3%)
		Classification	4
		Clustering	3
	Money laundering		1 (2.0%)
		Classification	1
			8
Insurance fraud	Crop Insurance fraud		2 (4.1%)
		Regression	2
	Healthcare insurance fraud		5 (10.2%)
		Classification	3
		Outlier Detection	1
		Visualization	1
	Automobile insurance fraud		17 (34.7%)
		Classification	11
		Prediction	2
		Regression	4
			24
Other related financial fraud	Corporate fraud		17 (34.7%)
		Classification	11
		Clustering	1
		Prediction	1
		Regression	4
			17
Total		49 (100%)	49

Table 7
Statistics of articles on financial fraud and data mining techniques.

No.	Techniques	Bank fraud		Insurance fraud			Other related financial fraud	Total
		Credit card fraud	Money laundering	Crop insurance fraud	Healthcare insurance fraud	Automobile insurance fraud	Corporate fraud	
1	Logistic model	1		1		9	5	16
2	Neural networks	2				2	6	10
3	Bayesian belief network	1				2	2	5
4	Decision trees	2				1	2	5
5	Naïve Bayes	1				2	1	4
6	Evolutionary algorithms	1				1	1	3
7	K-nearest neighbor	1				1	1	3
8	Probit model			1		2		3
9	Self-organizing map	1			1	1		3
10	Support vector Machine	1				1	1	3
11	CART	1					1	2
12	Discriminant analysis	2						2
13	Fuzzy logic					1	1	2
14	RIPPER	1					1	2
15	Ada boost algorithm	1						1
16	Association rule				1			1
17	Discounting learning algorithm				1			1
18	Hidden Markov Model	1						1
19	Multicriteria decision aid (MCDA)						1	1
20	Network analysis		1					1
21	Polymorphous (M-of-N) logic				1			1
22	Principal component analysis of RIDIT (PRIDIT)					1		1
23	Stacking variant methodology						1	1
24	UTilite's Additives DIScriminantes (UTADIS)						1	1
25	Visualization				1			1
26	Yield-switching model			1				1
	Total	17	1	3	5	24	25	75

FFD research studies is lacking. In this study, we conduct an extensive review of academic articles and provide a comprehensive bibliography and classification framework for the applications of data mining to FFD. Our intention is to inform both academics and practitioners of the areas in which specific data mining techniques can be applied to FFD, and to report and compile a systematic review of the burgeoning literature on FFD. Although our study cannot claim to be exhaustive, we believe that it will prove a useful resource for anyone interested in FFD research, and will help simulate further interest in the field.

The results of our study lead to the following conclusions.

- Of the four FF-based categories, Insurance fraud has attracted the greatest attention from researchers. Phua et al. [58] point out that insurance fraud is more likely to be committed by offenders, which may be why this type of fraud has gained so much research attention. Insurance fraud is also the area of FFD to which data mining techniques are most commonly applied (24 articles out of 49, or 49%), with automobile insurance fraud in particular being described in 17 out of the 24 articles. Artís et al. [5] argue that this is a subject of major concern for both companies and consumers.
- There are only a few studies on money laundering, mortgage fraud, mass marketing fraud, and securities and commodities fraud. Further, there is only one article that discusses the application of data mining to the detection of money laundering, and no articles reporting its application to the other three fraud types. Nevertheless, these fraudulent activities are important and deserve more research. Gao and Ye [36] emphasize that anti-money laundering research is of critical significance to national financial stability and international security, and the UN Office on Drugs and Crime (UNODC) estimates that the total amount of "black" money circulating worldwide reached 320 billion dollars in 2008 [69].
- The data mining techniques of outlier detection and visualization have seen only limited use. The lack of research on the application of outlier detection techniques to FFD may be due to the difficulty of detecting outliers. Indeed, Agyemang et al. [2] point out that outlier detection is a very complex task akin to finding a needle in a haystack. Distinct from other data mining techniques, outlier detection techniques are dedicated to finding rare patterns associated with very few data objects. In the field of FFD, outlier detection is highly suitable for distinguishing fraudulent data from authentic data, and thus deserves more investigation. Similarly, visualization techniques have a strong ability to recognize and present data anomalies, which could make the identification and quantification of fraud schemes much easier [64].

We suggest that one of the reasons for the limited number of relevant journal articles (49) published between 1997 and 2008 is the difficulty of obtaining sufficient research data. Fanning and Cogger [31] highlight the challenge of obtaining fraudulent financial

Table 8
Classification of articles by the categories of financial fraud and publication year.

FF-based categories	Fraudulent activities	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Total
Bank fraud	Credit card fraud	1		1							2		3	7
	Money laundering											1		1
Insurance fraud	Crop insurance fraud									1	1			2
	Healthcare insurance fraud	1				1	1		1		1			5
	Automobile insurance fraud	1	2	1	1		5		1	3		2	1	17
Other related financial fraud	Corporate fraud	3	3	1	1		3	1			1	1	3	17
Total		6	5	3	2	1	9	1	2	4	5	4	7	49

Table 9
Distribution of articles by journal title.

Journal title	Number	Percentage (%)
Journal of Risk and Insurance	8	16.3
Expert Systems with Applications	6	12.2
Managerial Auditing Journal	4	8.2
International Journal of Intelligent Systems in Accounting, Finance and Management	3	6.1
Auditing: A Journal of Practice & Theory	3	6.1
Insurance: Mathematics and Economics	2	4.1
American Journal of Agriculture Economics	1	2.0
Applied Economics	1	2.4
Computer Fraud and Security	1	2.4
Data Mining and Knowledge Discovery	1	2.4
Decision Support Systems	1	2.4
European Accounting Review	1	2.4
European Journal of Operational Research	1	2.4
Geneva Papers on Risk and Insurance	1	2.4
IEEE Intelligent Systems	1	2.4
IEEE Transactions on Dependable and Secure Computing	1	2.4
IEEE Transactions on Evolutionary Computation	1	2.4
IEEE Transactions on Knowledge and Data Engineering	1	2.4
IEEE Transactions on Neural Networks	1	2.4
Information and Security	1	2.4
International Journal of Computational Intelligence	1	2.4
International Journal of Information Technology and Decision Making	1	2.4
International Journal of Management	1	2.4
International Journal of Pattern Recognition and Artificial Intelligence	1	2.4
Journal of Law and Economics	1	2.4
Journal of Money Laundering Control	1	2.4
Managerial Finance	1	2.4
Risques	1	2.4
Topics in Health Information Management	1	2.4
Total	49	100

statements, and note that this creates enormous obstacles in FFD research. The most urgent challenge facing FFD is to bridge the gap between practitioners and researchers. The existing FFD research concentrates on particular types of data mining techniques or models, but future research should direct its attention toward finding more practical principles and solutions for practitioners to help them to design, develop, and implement data mining and business intelligence systems that can be applied to FFD.

We predict that increasing amounts of privacy-preserving financial data will be publicly available in the near future due to increased collaboration between practitioners and researchers, and that this should lead to more investigations of data mining techniques that can be applied to privacy-preserving data.

A further problem faced by FFD is that of cost sensitivity. The cost of misclassification (false positive and false negative errors) differs, with a false negative error (misclassifying a fraudulent activity as a normal activity) usually being more costly than a false positive error (misclassifying a normal activity as a fraudulent activity) [58]. Few studies have explicitly included cost in their FFD modeling [74], but future research on the application of data mining techniques to FFD problems should take into account cost sensitivity considerations.

This study has two major limitations. First, our review applied several keywords to search only nine online databases for articles published between 1997 and 2008. A future review could be expanded in scope. Second, we considered only articles written in English. Future research could be expanded to include relevant articles published in other languages.

Acknowledgements

The authors gratefully acknowledge the associate editor and reviewers' constructive comments on an earlier version of the paper. This research was partly supported by the National Natural Science Foundation of China (NSFC, project no.: 70801020), the Science and Technology Planning Project of Guangdong Province, China(project no.: 2010B010600034) and The Hong Kong Polytechnic University under a research grant number G-YX71 and the “211 Project” of Guangdong University of Foreign Studies.

References

[1] A. Agresti, Categorical Data Analysis, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1990.
[2] M. Agyemang, K. Barker, R. Alhajj, A comprehensive survey of numeric and symbolic outlier mining techniques, Intelligent Data Analysis 10 (6) (2006) 521–538.
[3] S.R. Ahmed, Applications of data mining in retail business, International Conference on Information Technology: Coding and Computing 2 (2) (2004) 455–459.
[4] M. Artís, M. Ayuso, M. Guillén, Modelling different types of automobile insurance fraud behaviour in the Spanish market, insurance, Mathematics and Economics 24 (1) (1999) 67–81.
[5] M. Artís, M. Ayuso, M. Guillén, Detection of automobile insurance fraud with discrete choice models and misclassified claims, The Journal of Risk and Insurance 69 (3) (2002) 325–340.
[6] J.A. Atwood, J.F. Robinson-Cox, S. Shaik, Estimating the prevalence and cost of yield-switching fraud in the federal crop insurance program, American Journal of Agricultural Economics 88 (2) (2006) 365–381.
[7] B. Bai, J. Yen, X. Yang, False financial statements: characteristics of China's listed companies and CART detecting approach, International Journal of Information Technology & Decision Making 7 (2) (2008) 339–359.
[8] BBC News, http://news.bbc.co.uk/1/hi/business/6636005.stm
[9] E.B. Belhadji, G. Dionne, F. Tarkhani, A model for the detection of insurance fraud, The Geneva Papers on Risk and Insurance 25 (4) (2000) 517–538.
[10] T.B. Bell, J.V. Carcello, A decision aid for assessing the likelihood of fraudulent financial reporting, Auditing: A Journal of Practice & Theory 19 (1) (2000) 169–174.
[11] L. Bermúdez, J.M. Pérez, M. Ayuso, E. Gómez, F.J. Vázquez, A. Bayesian Dichotomous, Model with asymmetric link for fraud in insurance, Insurance: Mathematics and Economics 42 (2) (2008) 779–786.
[12] M.J.A. Berry, G.S. Linoff, Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management, Second ed.Wiley, New York, 2004.
[13] R.J. Bolton, D.J. Hand, Statistical fraud detection: a review, Statistical Science 17 (3) (2002) 235–255.
[14] I. Bose, R.K. Mahapatra, Business data mining — a machine learning perspective, Information Management 39 (3) (2001) 211–225.
[15] P.L. Brockett, X. Xia, R.A. Derrig, Using Kononen's self-organizing feature map to uncover automobile bodily injury claims fraud, The Journal of Risk and Insurance 65 (2) (1998) 245–274.
[16] P.L. Brockett, R.A. Derrig, L.L. Golden, Fraud classification using principal component analysis of RIDITS, The Journal of Risk and Insurance 69 (3) (2002) 341–371.
[17] S.B. Caudill, M. Ayuso, M. Guillén, Fraud detection using a multinominal logit model with missing information, The Journal of Risk and Insurance 72 (4) (2005) 539–550.
[18] M.J. Cerullo, V. Cerullo, Using neural networks to predict financial reporting fraud, Computer Fraud & Security May/June (1999) 14–17.
[19] P.K. Chan, W. Fan, A.L. Prodromidir, S.L. Stalfo, Distributed data mining in credit card fraud detection, IEEE Intelligent Systems Nov/Dec (1999) 67–74.
[20] R. Chen, T. Chen, C. Lin, A new binary support vector system for increasing detection rate of credit card fraud, International Journal of Pattern Recognition and Artificial Intelligence 20 (2) (2006) 227–239.
[21] Coalition against Insurance Fraud, “Learn about fraud,” http://www.insurance-fraud.org/learn_about_fraud.htm.
[22] CULS, Cornell University Law School, White-Collar Crime: an overview, http://topics.law.cornell.edu/wex/White-collar_crime (2009)
[23] K.J. Crocker, S. Tennyson, Insurance fraud and optimal claims settlement strategies, Journal of Law and Economics 45 (2002) 469–507.
[24] R.A. Derrig, Insurance fraud, The Journal of Risk and Insurance 69 (3) (2002) 271–287.
[25] A. Deshmukh, J. Romine, P.H. Siegel, Measurement and combination of red flags to assess the risk of management fraud: a fuzzy set approach, Managerial Finance 23 (6) (1997) 35–48.
[26] A. Deshmukh, L. Talluru, A rule-based fuzzy reasoning system for assessing the risk of management fraud, International Journal of Intelligent Systems in Accounting, Finance & Management 7 (4) (1998) 223–241.
[27] J.R. Dorronsoro, F. Ginel, C. Sánchez, C.S. Cruz, Neural fraud detection in credit card operations, IEEE Transactions on Neural Networks 8 (4) (1997) 827–834.
[28] R.O. Duda, P.E. Hart, E.G. Stock, Pattern Classification, Wiley, New York, 2001.
[29] S.G. Eick, D.E. Fyock, Visualizing corporate data, AT&T Technical Journal 75 (1) (1996) 74–86.
[30] M. Eining, D.R. Jones, J.K. Loebbecke, Reliance on decision aids: an examination of auditors' assessment of management fraud, Auditing: A Journal of Practice & Theory 16 (2) (1997) 1–19.

[31] K.M. Fanning, K.O. Cogger, Neural network detection of management fraud using published financial data, International Journal of Intelligent Systems in Accounting, Finance & Management 7 (1) (1998) 21–41.
[32] T. Fawcett, F. Provost, Adaptive fraud detection, Data Mining and Knowledge Discovery 1 (3) (1997) 291–316.
[33] FBI, Federal Bureau of Investigation, Financial Crimes Report to the Public Fiscal Year, Department of Justice, United States, 2007, http://www.fbi.gov/publications/financial/fcs_report2007/financial_crime_2007.htm.
[34] FBI, Federal Bureau of Investigation New York Division, Department of Justice, United States, 2008, http://newyork.fbi.gov/dojpressrel/pressrel08/nyfo121108.htm.
[35] W.J. Frawley, G. Piatetsky-Shapiro, C.J. Matheus, Knowledge discovery in databases: an overview, AI Magazine 13 (3) (1992) 57–70.
[36] Z. Gao, M. Ye, A framework for data mining-based anti-money laundering research, Journal of Money Laundering Control 10 (2) (2007) 170–179.
[37] S. Ghosh, D.L. Reilly, Credit card fraud detection with a neural-network, 27th Annual Hawaii International, Conference on System Science 3 (1994) 621–630.
[38] P. Green, J.H. Choi, Assessing the risk of management fraud through neural network technology, Auditing: A Journal of Practice & Theory 16 (1) (1997) 14–28.
[39] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Second ed, Morgan Kaufmann Publishers, 2006, pp. 285–464.
[40] M. Haskett, An Introduction to Data Mining, Part 2, Analyzing the Tools and Techniques, Enterprise System Journal, 2000.
[41] H. He, J. Wang, W. Graco, S. Hawkins, Application of neural networks to detection of medical fraud, Expert Systems with Applications 13 (4) (1997) 329–336.
[42] C. Holton, Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multi-billion dollar problem, Decision Support Systems 46 (4) (2009) 853–864.
[43] Y. Jin, R.M. Rejesus, B.B. Little, Binary choice models for rare events data: a crop insurance fraud application, Applied Economics 37 (7) (2005) 841–848.
[44] J.L. Kaminski, Insurance Fraud, OLR Research Report, http://www.cga.ct.gov/2005/rpt/2005-R-0025.htm. 2004
[45] E. Kirkos, C. Spathis, Y. Manolopoulos, Data mining techniques for the detection of fraudulent financial statements, Expert Systems with Applications 32 (4) (2007) 995–1003.
[46] S. Kotsiantis, E. Koumanakos, D. Tzelepis, V. Tampakas, Forecasting fraudulent financial statements using data mining, International Journal of Computational Intelligence 3 (2) (2006) 104–110.
[47] Y. Kou, C. Lu, S. Sirwongwattana, Y. Huang, Survey of fraud detection techniques, IEEE International Conference on Networking, Sensing & Control (2004) 749–754.
[48] W. Lee, S. Stolfo, Data Mining Approaches for Intrusion Detection, 7th USENIX Security Symposium, San Antonio, TX, 1998.
[49] J. Li, K. Huang, J. Jin, J. Shi, A survey on statistical methods for health care fraud detection, Health Care Management Science 11 (3) (2008) 275–287.
[50] J.W. Lin, M.I. Hwang, J.D. Becker, A fuzzy neural network for assessing the risk of fraudulent financial reporting, Managerial Auditing Journal 18 (8) (2003) 657–665.
[51] J.A. Major, D.R. Riedinger, EFD: a hybrid knowledge/statistical-based system for the detection of fraud, The Journal of Risk and Insurance 69 (3) (2002) 309–324.
[52] S. Mitra, S.K. Pal, P. Mitra, Data mining in soft computing framework: a survey, IEEE Transactions on Neural Networks 13 (1) (2002) 3–14.
[53] E.W.T. Ngai, L. Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: a literature review and classification, Expert Systems with Applications 36 (2) (2009) 2592–2602.
[54] S. Owusu-Ansah, G.D. Moyes, P.B. Oyelere, P. Hay, An empirical analysis of the likelihood of detecting fraud in New Zealand, Managerial Auditing Journal 17 (4) (2002) 192–204.
[55] Oxford Concise English Dictionary, Tenth ed, Publisher, 1999.
[56] J. Pathak, N. Vidyarthi, S.L. Summers, A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims, Managerial Auditing Journal 20 (6) (2005) 632–644.
[57] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, 1988.
[58] C. Phua, V. Lee, K. Smith, R. Gayler, A comprehensive survey of data mining-based fraud detection research, Artificial Intelligence Review (2005) 1–14.
[59] J. Pinquet, M. Ayuso, M. Guillén, Selection bias and auditing policies for insurance claims, The Journal of Risk and Insurance 74 (2) (2007) 425–440.
[60] J.T.S. Quah, M. Sriganesh, Real-time credit card fraud detection using computational intelligence, Expert Systems with Applications 35 (4) (2008) 1721–1732.
[61] D. Sánchez, M.A. Vila, L. Cerda, J.M. Serrano, Association rules applied to credit card fraud detection, Expert Systems with Applications 36 (2) (2009) 3630–3640.
[62] S. Sharma, Applied Multivariate Techniques, Wiley, New York, 1996.
[63] M.J. Shaw, C. Subramaniam, G.W. Tan, M.E. Welge, Knowledge management and data mining for marketing, Decision Support System 31 (1) (2001) 127–137.
[64] L. Sokol, B. Garcia, J. Rodriguez, M. West, K. Johnson, Using data mining to find fraud in HCFA health care claims, Topics in Health Information Management 22 (1) (2001) 1–13.
[65] C.T. Spathis, Detecting false financial statements using published data: some evidence from Greece, Managerial Auditing Journal 17 (4) (2002) 179–191.
[66] C.T. Spathis, M. Doumpos, C. Zopounidis, Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques, The European Accounting Review 11 (3) (2002) 509–535.
[67] A. Srivastava, A. Kundu, S. Sural, A.K. Majumdar, Credit card fraud detection using hidden Markov model, IEEE Transactions on Dependable and Secure Computing 5 (1) (2008) 37–48.
[68] M. Sternberg, R.G. Reynolds, Using cultural algorithms to support re-engineering of rule-based expert systems, in dynamic performance environments: a case study in fraud detection, IEEE Transactions on Evolutionary Computation 1 (4) (1997) 225–243.
[69] stopthedrugwar.org, http://stopthedrugwar.org/chronicle/570/costa_UNODC_-drug_trade_banks, 30 Jan. 2009
[70] M. Syeda, Y. Zhang, Y. Pan, Parallel granular neural networks for fast credit card fraud detection, 2002, IEEE International Conference on Fuzzy Systems 1 (2002) 572–577.
[71] P. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, First ed.Addison-Wesley Longman Publishing Co., Inc, 2005.
[72] S. Tennyson, P. Salsas-Forn, Claims auditing in automobile insurance: fraud detection and deterrence objectives, The Journal of Risk and Insurance 69 (3) (2002) 289–308.
[73] E. Turban, J.E. Aronson, T.P. Liang, R. Sharda, Decision Support and Business Intelligence Systems, Eighth ed, Pearson Education, 2007.
[74] S. Viaene, M. Ayuso, M. Guillén, D. Van Gheel, G. Dedene, Strategies for detecting fraudulent claims in the automobile insurance industry, European Journal of Operational Research 176 (1) (2007) 565–583.
[75] S. Viaene, G. Dedene, R.A. Derrig, Auto claim fraud detection using bayesian learning neural networks, Expert Systems with Applications 29 (3) (2005) 653–666.
[76] S. Viaene, R.A. Derrig, B. Baesens, G. Dedene, A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection, The Journal of Risk and Insurance 69 (3) (2002) 373–421.
[77] S. Viaene, R.A. Derrig, G. Dedene, A case study of applying boosting naive Bayes to claim fraud diagnosis, IEEE Transactions on Knowledge and Data Engineering 16 (5) (2004) 612–620.
[78] J. Wang, Y. Liao, T. Tsai, G. Hung, Technology-based financial frauds in Taiwan: issue and approaches, IEEE Conference on: Systems, Man and Cyberspace Oct (2006) 1120–1124.
[79] A. Webb, Statistical Pattern Recognition, Arnold, London, 1999.
[80] H.I. Weisberg, R.A. Derrig, Quantitative methods for detecting fraudulent automobile bodily injury claims, Risques 35 (1998) 75–101.
[81] J. Welch, T.E. Reeves, S.T. Welch, Using a genetic algorithm-based classifier system for modeling auditor decision behavior in a fraud setting, International Journal of Intelligent Systems in Accounting, Finance & Management 7 (3) (1998) 173–186.
[82] K. Yamanishi, J. Takeuchi, G. Williams, P. Milne, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, Data Mining and Knowledge Discovery 8 (3) (2004) 275–300.
[83] W. Yang, S. Hwang, A process-mining framework for the detection of healthcare fraud and abuse, Expert Systems with Applications 31 (1) (2006) 56–68.
[84] I. Yeh, C. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Applications 36 (2) (2008) 2473–2480.
[85] J. Yuan, C. Yuan, X. Deng, C. Yuan, The effects of manager compensation and market competition on financial fraud in public companies: an empirical study in China, International Journal of Management 25 (2) (2008) 322–335.
[86] Yue, X. Wu, Y. Wang, Y. Li, C. Chu, A review of data mining-based financial fraud detection research, international conference on wireless communications Sep, Networking and Mobile Computing (2007) 5519–5522.
[87] V. Zaslavsky, A. Strizhak, Credit card fraud detection using self-organizing maps, Information & Security 18 (2006) 48–63.
[88] D. Zhang, L. Zhou, Discovering golden nuggets: data mining in financial application, IEEE Transactions on Systems, Man and Cybernetics 34 (4) (2004) Nov.

Prof. Eric Ngai is a Professor in the Department of Management and Marketing at The Hong Kong Polytechnic University. His current research interests are in the areas of E-commerce, Supply Chain Management, Decision Support Systems and RFID Technology and Applications. He has published papers in a number of international journals including *MIS Quarterly, Journal of Operations Management, Decision Support Systems, IEEE Transactions on Systems, Man and Cybernetics, Information & Management, Production & Operations Management,* and others. He is an Associate Editor of *European Journal of Information Systems* and serves on editorial board of six international journals. Prof. Ngai has attained an *h*-index of 13, and received 510 citations, *ISI Web of Science*.

Dr. Yong Hu is currently an Associate Professor and Chair in the Department of E-commerce, and Director of Institute of Business Intelligence and Knowledge Discovery at the Guangdong University of Foreign Studies. He received his B.Sc in Computer Science, M.Phil and Ph.D. in Management Information Systems from Sun Yat-Sen University. His research interests are in the areas of business intelligence, software project risk management, e-commerce and decision support systems. He has published in a number of journals and conferences such as DSS, ESWA, JECO and IEEE ICDM. Dr. Hu's research is supported by the National Natural Science Foundation, the Science and Technology Planning Project of Guangdong Province, and “211 Project” of the Guangdong University of Foreign Studies.

Dr. Y. H. Wong is associate professor, Department of Management and Marketing, The Hong Kong Polytechnic University. His publications include 3 books, *Guanxi*: Relationship Marketing in a Chinese Context, Handbook of Research on Ubiquitous Commerce for Creating the Personalized Marketplace, Financial Planning and Wealth Management and refereed journal articles, such as, *Industrial Marketing Management, International Business Review, European Journal of Marketing, Journal of Services Marketing and Journal of Business Ethics*, etc.

Yi-Jun Chen is an M.Phil student in Department of Computing and Decision Science, Lingnan University, Hong Kong S.A.R of China. He received his BSc degree in information and computational science from Sun Yat-Sen University. His research interest includes data mining, Bayesian networks and business intelligence.

Xin Sun is an M.Phil student in Guangdong University of Foreign Studies and working as an assistant researcher in Institute of Business Intelligence and Knowledge Discovery. She has received her BSc in Mathematics and Applied Mathematics from Sun Yat-Sen University. Her research interest is software project risk management and business intelligence.