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

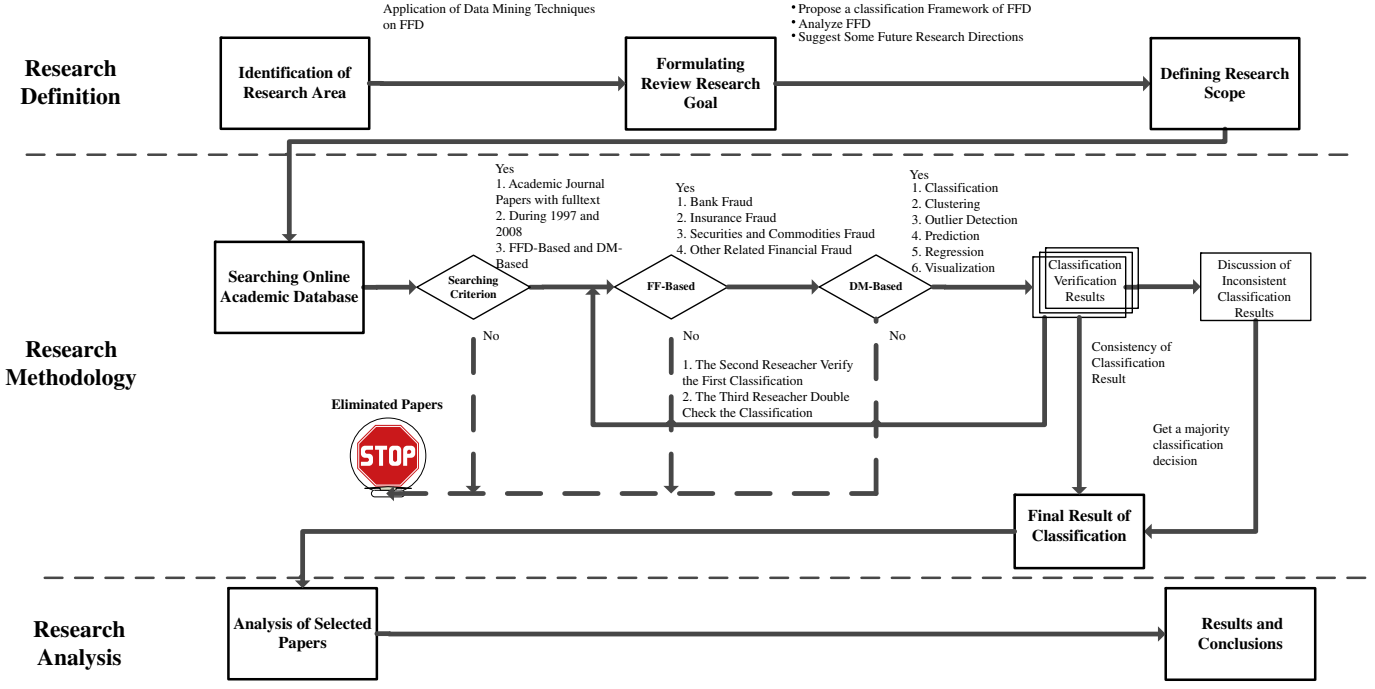

Fig. 1. Methodological framework for research.