Expert Systems with Applications 41 (2014) 4360–4372

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Topological pattern discovery and feature extraction for fraudulent financial reporting

Shin-Ying Huang [a,*], Rua-Huan Tsaih [b], Fang Yu [b]

[a] Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan
[b] Department of Management Information Systems, National Chengchi University, Taipei 11605, Taiwan

ARTICLE INFO

Keywords:
Unsupervised learning
Growing Hierarchical Self-Organizing Map
Data mining
Fraudulent financial reporting

ABSTRACT

Fraudulent financial reporting (FFR) involves conscious efforts to mislead others regarding the financial condition of a business. It usually consists of deliberate actions to deceive regulators, investors or the general public that also hinder systematic approaches from effective detection. The challenge comes from distinguishing dichotomous samples that have their major attributes falling in the same distribution. This study pioneers a novel dual GHSOM (Growing Hierarchical Self-Organizing Map) approach to discover the topological patterns of FFR, achieving effective FFR detection and feature extraction. Specifically, the proposed approach uses fraudulent samples and non-fraudulent samples to train a pair of dual GHSOMs under the same training parameters and examines the hypotheses for counterpart relationships among their subgroups taking advantage of unsupervised learning nature and growing hierarchical structures from GHSOMs. This study further presents (1) an effective classification rule to detect FFR based on the topological patterns and (2) an expert-competitive feature extraction mechanism to capture the salient characteristics of fraud behaviors. The experimental results against 762 annual financial statements from 144 public-traded companies in Taiwan (out of which 72 are fraudulent and 72 are non-fraudulent) reveal that the topological pattern of FFR follows the non-fraud-central spatial relationship, as well as shows the promise of using the topological patterns for FFR detection and feature extraction.

1. Introduction

Capital providers, e.g., the creditors, rely on accounting-based numbers as a persistent and traditional standard for assessment of a firm's financial soundness and viability. However, fraudulent financial reporting (FFR) still exists as a type of financial fraud that involves the intentional misstatement or omission of material information from organizations' financial reports (Beasley, Carcello, & Hermanson, 1999). FFR can lead not only to significant risks for creditors and stockholders but also to financial crises for the capital market (Association of Certified Fraud Examiners (ACFE), 2008). All too often, investors and creditors encounter situations in which it appears that the auditors and/or the audit committees were not effective (ACFE, 2008; Beasley et al., 1999). The study speculates that being ineffective to detect FFR may be due to the unusual topological pattern of the financial statements.

FFR is the deliberate action of issuing misleading financial statements in an effort to avoid negative opinions about the financial stability of a particular business or other type of institution. It is important to note that FFR takes place when there is a conscious effort to mislead others regarding the financial condition of a business or other entity. This type of reckless conduct may involve omitting relevant data from the report, or even altering figures as a means of deceiving regulators, investors and consumers in general. Rather than some data being overlooked by accident, the intentional omissions are carefully chosen so as to alter the overall image created by the financial reports that are issued to investors and ultimately to the general public. Typically, supporting documents are altered as part of the FFR, in an effort to support the false impression. This additional deception only serves to increase the level of duplicity involved. By contrast, an unintentional omission can often be discovered by reading the content of the supporting documents to identify what information was overlooked.

The aforementioned statements imply that the fraudulent financial statements are manipulated to be plausibly similar to some non-fraudulent ones such that the fraudulent ones can be overlooked and treated as the healthy one. That is, in terms of the high-dimensional space consisted of financial ratios derived from fraudulent and non-fraudulent financial statements, the fraudulent data tend to locate at the neighborhood of their non-fraudulent counterparts. This leads to the topological assumption

* Corresponding author. Address: No. 128, Academia Road, Section 2, Nankang, Taipei 11529, Taiwan.
E-mail addresses: smichelle19@gmail.com, smichelle19@citi.sinica.edu.tw (S.-Y. Huang).

http://dx.doi.org/10.1016/j.eswa.2014.01.012
0957-4174/

regarding the high-dimensional financial ratios derived from FFR data: the spatial distributions of fraudulent and non-fraudulent data are similar and satisfy the following *non-fraud-central spatial hypothesis*:

The non-fraud-central spatial hypothesis: As shown in Fig. 1, the non-fraudulent data locate in segregated clusters; within each cluster, the fraudulent counterparts locate around the outer circle.

Previous FFR-related research focused mainly on the nature of FFR (ACFE, 2008; Association of Certified Fraud Examiners (ACFE), 2013; Beasley et al., 1999; Green & Choi, 1997; La Porta, Lopez-de-Silanes, Shleifer, & Vishny, 1999) or the prediction of FFR (Dechow, Ge, Larson, & Sloan, 2007; Fanning & Cogger, 1998; Ngai, Yong, Wong, Chen, & Sun, 2011; Pulakkazhy & Balan, 2013; Sharma & Panigrahi, 2012). Research on the nature of FFR often uses the case study approach and provides a descriptive analysis of the characteristics of FFR and commonly it is used techniques. The Association of Certified Fraud Examiners (Association of Certified Fraud Examiners (ACFE), 2008) analyzed the nature of occupational fraud schemes and emphasized the importance of adequate internal control mechanisms. The manual of ACFE describes hundreds of fraud schemes, as well as providing information about basic accounting concepts and managers' and auditors' responsibilities to detect fraud, and also explores why people commit fraud and what can be done to prevent it. Topics include crime causation, white-collar crime, occupational fraud, fraud prevention, fraud risk assessment, and the ACFE Code of Professional Ethics (Association of Certified Fraud Examiners (ACFE), 2013). Beasley et al. (1999) stated that there are three basic categories of FFR: improper revenue recognition, overstatement of assets, and others. Improper revenue recognition includes recording fictitious revenues (FC1), recording revenues prematurely (FC2), and no description/overstated revenues (FC3). Overstatement of assets includes overstating existing assets (FC4), recording fictitious assets or assets not owned (FC5), and capitalizing items that should be expensed (FC6). The others category includes understatement of expenses/liabilities (FC7), misappropriation of assets (FC8), inappropriate disclosure (FC9), and other miscellaneous techniques (FC10).

Green and Choi (1997) applied the back-propagation neural network for FFR detection with a model that used five ratios and three accounts as inputs, and the result showed the back-propagation neural network has significant potential ability as a fraud detection tool. Fanning and Cogger (1998) proposed a generalized adaptive neural network algorithm called AutoNet to detect FFR and compared their model against linear and quadratic discriminant analysis and logistic regression, concluding that AutoNet is more effective in detecting fraud than standard statistical methods. Classification techniques based on neural network, regression and decision tree are used for classifying fraudulent ratios in the statements from the non-fraudulent data (Sharma & Panigrahi, 2012).

The challenge of FFR detection is that the fraudulent statements may contain overstated assets, sales and profits or it may understate losses and liabilities. Even though these statements may have been audited, these kinds of frauds are hard to be detected using normal auditing procedures. Pulakkazhy and Balan (2013) stated that, with the help of data mining method, the detected patterns help the bank to forecast future events and do decision-making. However, to the best of our knowledge, none of the related studies addresses the spatial relationships among financial ratios derived from fraudulent and non-fraudulent samples. As shown in our experiments, such topological relationship may render conventional classification algorithms and tools unable to successfully separate fraudulent and non-fraudulent data.

In both theory and practice, there is a lack of means of systematically discovering the spatial relationships of such dichotomous (e.g., fraudulent and non-fraudulent) samples, particularly for those whose input space has high dimensionality. To help explore the nature of FFR and predict the occurrence of FFR, we need an analytic approach which is capable of systematically discovering the spatial relationships of dichotomous samples.

Based on Growing Hierarchical Self-Organizing Map (GHSOM), an extension of Self-Organizing Map (SOM) which is an unsupervised neural network for clustering and has been applied for the problems in banking domain (Carlos, 1996; Shih, 2011), we extend its property of topology preserving to systematically discover the spatial relationships of dichotomous samples. That is, for the pattern recognition of dichotomous data, the dual GHSOM approach is proposed to examine the non-fraud-central spatial hypothesis and the following fraud-central spatial hypothesis via a pair of dual GHSOMs.

This study proposes a novel application of unsupervised Neural Networks for effective pattern recognition of dichotomous (e.g., fraudulent and non-fraudulent) data. For a general pattern recognition of dichotomous data, the *dual GHSOM* approach is proposed to examine the *non-fraud-central spatial hypothesis* and the following *fraud-central spatial hypothesis* via a pair of dual GHSOMs.

The fraud-central spatial hypothesis: As shown in Fig. 2, the fraudulent data locate in segregated clusters; within each cluster, the non-fraudulent counterparts locate around the outer circle.

Specifically, the proposed approach uses fraudulent data and non-fraudulent data to train a pair of GHSOMs respectively under the same training parameter settings, and then examines the topological pattern among their counterpart subgroups. That is, the proposed approach consists of two main mechanisms: the classification mechanism and the feature-extraction mechanism. In the classification mechanism, the non-fraudulent and fraudulent

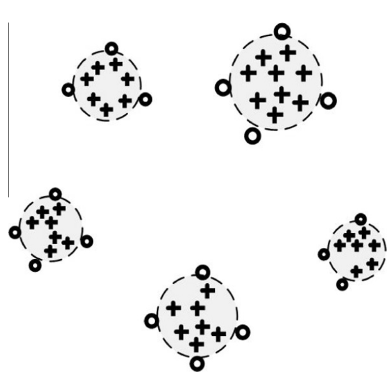

Fig. 1. The non-fraudulent data (+s) locate in segregated clusters; within each cluster, the fraudulent counterparts (os) locate around the outer circle.

Fig. 2. The fraudulent data (os) locate in segregated clusters; within each cluster, the non-fraudulent counterparts (+s) locate around the outer circle.