國立政治大學資訊管理研究所

博士學位論文

指導教授：蔡瑞煌博士

適用於財務舞弊偵測之決策支援系統的對偶方法

A Dual Approach for Decision Support in Financial Fraud Detection

研究生：黃馨瑩

中華民國 101 年 7 月

誌 謝

博士班五年的生涯，能夠有幸走到畢業的這一刻，我感覺到有點不真實，因為學術生涯是一條無止境的道路，不單單只是通過了這個關卡，背後其實有好多的故事和人生經驗的累積，也許向前是淨土、或沙漠，我都要用充滿感謝的心情，謝謝好多人，謝謝我所經歷的人生歷練，讓我有今天的模樣，也許還不成熟，但至少邁出了人生的一小步。

首先，感謝我的家人，爸爸和媽媽是我的精神支柱，博班漫長的道路上，各式的挑戰與歷練他們都感同身受，他們的關心和傾聽鼓勵了我讓我有更堅定的信念向前邁進，對他們的感謝難以言語，最大的心願是能讓他們快樂安心、沒有後顧之憂。還有，我要感謝在天上的奶奶，和特地從屏東北上參加我畢業典禮的外婆，奶奶在我考上博班時，包了一個紅包給我，裡面有一萬塊，我本來想說通過資格考時再拿，奶奶過世以後，我決定要等畢業再去拿這個紅包，奶奶的關心一直是鼓勵我的力量，每當我意志力薄弱時，我都會在心裡默默跟奶奶說話，然後我就恢復滿滿的力量繼續努力。每次回屏東時，外婆都會關心我的學業，八十幾歲了還特地坐高鐵北上參加我的畢業典禮，因此我覺得很感激。謝謝我的老弟、瀞芝和姪子宥涂，我們常常分享生活中的點滴，可愛的宥涂也讓生活充滿樂趣，希望老弟賢伉儷能幸福美滿，期待即將出生的妹妹讓整個家庭更加熱鬧喔。

再來，我要感謝我的指導老師蔡瑞煌老師，在您的指導下，我學到了一步一腳印的精神，您總是在我顯露出緊張時安撫我叫我不要擔心，還有常鼓勵我 keep going 繼續努力，您在學術專業上教我要有嚴謹的態度，回答別人的詢問時也要經過思考再回答，然後要主動積極，因為機會不會主動來找我。這些我都會謹記在心，並學習您做為一個學者的風範。感謝您無私的指導，讓我從撰寫研究期刊和參加研討會的過程中累積研究的能力；感謝您給與學生一定程度的自主性，藉此引導我學習如何獨立思考，設法發揮研究的貢獻。感謝求學的路上有緣向老師學習。還有，我要感謝會計系林宛瑩老師，林老師教導我很多會計上的專業知識，在與老師參與研討會的經驗中也學習到很多，親切的林老師每次見面都會鼓勵我，感謝林老師讓我有跨領域的思維和樂觀的態度。此外，感謝我的論文評審委員李書行老師、許瑋元老師、郁方老師，謝謝老師們指正論文應當修改的地方，並且給予專業上的建議和論文撰寫上的指教，讓我在口試的過程裡吸收了大量的知識，對往後的論文撰寫獲益良多。

要感謝的人還有好多，謝謝研究所指導老師邱志洲老師在我念博班時仍時常對我加油打氣，並嘗試給予我其他研究學習的機會讓我磨練研究的能力。謝謝大

學指導老師洪朝富老師的教導讓我能踏入資管研究領域並對方法論產生興趣，謝謝老師的關心與鼓勵。謝謝博班陳春龍老師在方法論上的指導，管郁君老師在導生聚會時的關心與叮嚀，以及謝謝李有仁老師給與我論文上的建議，和曾經修過課程的楊建民老師、苑守慈老師、曾淑峰老師、姜國輝老師、余千智老師、趙玉老師等。老師們認真的處事哲學與關心學生的態度都為我立下了好榜樣，是我學習的目標。

求學的路上，感謝有以下好友的相伴，分享生活與彼此打氣，讓生活增添了很多的色彩：

博士班同班同學泉錫大哥、亭妤姐、學隆、耀中、凱康、立人同學。珍惜與各位一同修課研究以及定期互相鼓勵的聚會，也希望大家在學術及事業方面皆順利。

博班學長姐宇瑞、國華、木花、筱芳、逸寧、燕豪、嘉仁、廣豐、正育、俊隆、清為、明汶

博班學弟喻翔、偉成、逢毅、沛宏、文進、緒浩、芳凱、瑋佑、志得、千翔

系辦認真的雨儒與詩晴，謝謝妳們幫忙我很多系上行政的事情，辛苦了！

研究所同伴育宏、正弘、鴻昌、阿寶、濬遠學長

高中好友小娟、佩如、思群、怡任、雞腿、你的頭、河童、豌豆、阿湯、鄉下、小晏、大炳、佩佩

大學同學玉婷、舒靖、琬鈴、馨慧、湘文、秉妤、有寧、Freddy、小明、思余、文育、小瓜呆、cp、凱瑟琳

大學天倫社的瑞雯、美方、純瑜、宜真、阿牛、玉婷學姊

國中同學阿孝、阿懋、黃舜、褚聰、熊、馨瑋、鍋貼、佳慧、瑞萼、德怡

如果有遺漏的，請當作我放在心裡了。

博班學業的完成，長遠看來其實是人生職業生涯的起點，儘管畢業，總覺得自己還不夠有資格，因此只能不斷地學習與歷練，讓自己更趨成熟穩健。真正令我感到歷久彌新的，是這將近五年間我所經歷的人生變化，從懵懂具有衝勁的博一生，到開始看長遠生涯的博四博五生。緣起、緣滅，也經歷過親友生、老、病、死的成長。升格當上了姑姑，也開始準備邁向三十歲後的人生…

學業在我的生涯裡一直與我並行，除了 learn how to learn，我學會了珍惜當下，珍惜與我有緣的親朋好友們，不管我們的緣分或長或短，我都記著了，不管時間經過多久，我們的情誼永遠長存，也希望大家都有幸福的人生、美滿的家庭、順利圓夢、築夢踏實、身體健康、平平安安、快快樂樂。

馨瑩　筆

Content

List of Tables

List of Figures

Abstract

The Growing Hierarchical Self-Organizing Map (GHSOM) is extended from the Self-Organizing Map (SOM). The GHSOM's unsupervised learning nature such as the adaptive group size as well as the hierarchy structure renders its availability to discover the statistical salient features from the clustered groups, and could be used to set up a classifier for distinguishing abnormal data from regular ones based on spatial relationships between them.

Therefore, this study utilizes the advantage of the GHSOM and pioneers a novel dual approach (i.e., a proposal of a DSS architecture) with two GHSOMs, which starts from identifying the counterparts within the clustered groups. Then, the classification rules are formed based on a certain spatial hypothesis, and a feature extraction mechanism is applied to extract features from the fraud clustered groups. The dominant classification rule is adapted to identify suspected samples, and the results of feature extraction mechanism are used to pinpoint their relevant input variables and potential fraud activities for further decision aid.

Specifically, for the financial fraud detection (FFD) domain, a non-fraud (fraud) GHSOM tree is constructed via clustering the non-fraud (fraud) samples, and a non-fraud-central (fraud-central) rule is then tuned via inputting all the training samples to determine the optimal discrimination boundary within each leaf node of the non-fraud (fraud) GHSOM tree. The optimization renders an adjustable and effective rule for classifying fraud and non-fraud samples. Following the implementation of the DSS architecture based on the proposed dual approach, the decision makers can objectively set their weightings of type I and type II errors. The classification rule that dominates another is adopted for analyzing samples. The dominance of the non-fraud-central rule leads to an implication that most of fraud samples cluster around the non-fraud counterpart, meanwhile the dominance of fraud-central rule leads to an implication that most of non-fraud samples cluster around the fraud counterpart.

Besides, a feature extraction mechanism is developed to uncover the regularity of input variables and fraud categories based on the training samples of each leaf node of a fraud GHSOM tree. The feature extraction mechanism involves extracting the variable features and fraud patterns to explore the characteristics of fraud samples within the same leaf node. Thus can help decision makers such as the capital providers

evaluate the integrity of the investigated samples, and facilitate further analysis to reach prudent credit decisions.

The experimental results of detecting fraudulent financial reporting (FFR), a sub-field of FFD, confirm the spatial relationship among fraud and non-fraud samples. The outcomes given by the implemented DSS architecture based on the proposed dual approach have better classification performance than the SVM, SOM+LDA, GHSOM+LDA, SOM, BPNN and DT methods, and therefore show its applicability to evaluate the reliability of the financial numbers based decisions. Besides, following the SOM theories, the extracted relevant input variables and the fraud categories from the GHSOM are applicable to all samples classified into the same leaf nodes. This principle makes that the extracted pre-warning signal can be applied to assess the reliability of the investigated samples and to form a knowledge base for further analysis to reach a prudent decision. The DSS architecture based on the proposed dual approach could be applied to other FFD scenarios that rely on financial numbers as a basis for decision making.

Keywords: Growing Hierarchical Self-Organizing Map; Unsupervised Neural Networks; Classification; Financial Fraud Detection; Fraudulent Financial Reporting.

摘　要

增長層級式自我組織映射網路(GHSOM)屬於一種非監督式類神經網路，為自我組織映射網路(SOM)的延伸，擅長於對樣本分群，以輔助分析樣本族群裡的共同特徵，並且可以透過族群間存在的空間關係假設來建立分類器，進而辨別出異常的資料。

因此本研究提出一個創新的對偶方法(即為一個建立決策支援系統架構的方法)分別對舞弊與非舞弊樣本分群，首先兩類別之群組會被配對，即辨識某一特定無弊群體的非舞弊群體對照組，針對這些配對族群，套用基於不同空間假設所設立的分類規則以檢測舞弊與非舞弊群體中是否有存在某種程度的空間關係，此外並對於舞弊樣本的分群結果加入特徵萃取機制。分類績效最好的分類規則會被用來偵測受測樣本是否有舞弊的嫌疑，萃取機制的結果則會用來標示有舞弊嫌疑之受測樣本的舞弊行為特徵以及相關的輸入變數，以做為後續的決策輔助。

更明確地說，本研究分別透過非舞弊樣本與舞弊樣本建立一個非舞弊 GHSOM 樹以及舞弊 GHSOM 樹，且針對每一對 GHSOM 群組建立分類規則，其相應的非舞弊/舞弊為中心規則會適應性地依循決策者的風險偏好最佳化調整規則界線，整體而言較優的規則會被決定為分類規則。非舞弊為中心的規則象徵絕大多數的舞弊樣本傾向分布於非舞弊樣本的周圍，而舞弊為中心的規則象徵絕大多數的非舞弊樣本傾向分布於舞弊樣本的周圍。

此外本研究加入了一個特徵萃取機制來發掘舞弊樣本分群結果中各群組之樣本資料的共同特質，其包含輸入變數的特徵以及舞弊行為模式，這些資訊將能輔助決策者(如資本提供者)評估受測樣本的誠實性，輔助決策者從分析結果裡做出更進一步的分析來達到審慎的信用決策。

本研究將所提出的方法套用至財報舞弊領域(屬於財務舞弊偵測的子領域)進行實證，實驗結果證實樣本之間存在特定的空間關係，且相較於其他方法如 SVM、SOM+LDA 和 GHSOM+LDA 皆具有更佳的分類績效。因此顯示本研究所提出的機制可輔助驗證財務相關數據的可靠性。此外，根據 SOM 的特質，即任何受測樣本歸類到某特定族群時，該族群訓練樣本的舞弊行為特徵將可以代表此受測樣本的特徵推論。這樣的原則可以用來協助判斷受測樣本的可靠性，並可供持續累積成一個舞弊知識庫，做為進一步分析以及制定相關信用決策的參考。本研究所提出之基於對偶方法的決策支援系統架構可以被套用到其他使用財務數據為資料來源的財務舞弊偵測情境中，作為輔助決策的基礎。

關鍵詞： 增長層級式自我組織映射網路；非監督式類神經網路；分類；財務舞弊偵測；財務報表舞弊

1. Introduction

This study proposes a dual approach as a Decision Support System (DSS) architecture based on the Growing Hierarchical Self-Organizing Map (GHSOM) (Dittenbach et al., 2000; Dittenbach et al., 2002; Rauber et al., 2002), a type of unsupervised artificial neural networks (ANN), for the decision support in financial fraud detection (FFD). FFD involves distinguishing fraudulent financial data from authentic data, disclosing fraudulent behavior or activities, and enabling decision makers to develop appropriate strategies to decrease the impact of fraud (Lu and Wang, 2010). The decision for FFD can be aided by statistical methods such as the logistic regression, as well as data mining tools such as the ANN, in which the ANN has been widely used and plays an important role in FFD (Lu and Wang, 2010). Among the ANN applications in FFD, the Self-Organizing Map (SOM) (Kohonen, 1982) has been adopted in diagnosing bankruptcy (Carlos, 1996). The major advantage of the SOM is its great visualization capability of topological relationship among the high-dimensional inputs in the low-dimensional view. Other advantages are adaptive (i.e., the clustering can be redone if new training samples are set) and robust (i.e., the pattern recognition ability). There are numerous applications involving the SOM and the most widespread use is the identification and visualization of natural groupings in the sample data sets. However, the weaknesses of the SOM include its predefined and fixed topology size and its inability to provide the hierarchical relations among samples (Dittenbach et al., 2000).

An improvement of the SOM has been done by Dittenbach, Merkl and Rauber (2000). They developed the GHSOM which addresses the issue of fixed network architecture of the SOM through developing a multilayer hierarchical network structure. The flexible and hierarchical feature of the GHSOM generates delicate clustered subgroups with heterogeneous features, and makes it a powerful and versatile data mining tool. The GHSOM has been used in many fields such as the image recognition, web mining, text mining, and data mining (Dittenbach et al., 2000; Schweighofer et al., 2001; Dittenbach et al., 2002; Rauber et al., 2002; Shih et al., 2008; Zhang and Dai, 2009; Tsaih et al., 2009). It is worth of knowing that the GHSOM can be a useful clustering tool to do the pre-processing of feature extraction for a certain application field.

In general, the GHSOM mainly takes the task of clustering and then visualizing the clustering results. To accomplish other purposes such as prediction or classification, the neural networks must be complemented with a statistical study of the available information (Serrano, 1996). However, this study finds that the development of the GHSOM into a classification model has been limited studied (Hsu et al., 2009; Lu and Wang, 2010; Guo et al., 2011). Besides, other than the hierarchical feature, the GHSOM studies have rarely provided the topological insight into high-dimensional inputs.

To better utilize the advantage of the GHSOM for the purpose of classification and feature extraction in helping FFD, this study pioneers a DSS architecture based on the proposed dual approach which helps extract the nature of the distinctive characteristics among different clustered groups generated by the GHSOM. This study develops an innovative way of observing the clustered data to form the optimal classification rule, and revealing more information regarding the relevant input variables and the potential fraud categories for the suspected samples as the knowledge base for facilitating FFD decision making.

This study examines the following topological relationships regarding high-dimensional inputs, of which there are two types: fraud and non-fraud, and matches the fraud counterpart of each non-fraud subgroup and vice versa. This study assumes that there is a certain spatial relationship among fraud and non-fraud samples. The spatial hypothesis: The spatial distributions of fraud samples and their non-fraud counterparts are identical, and the spatial distributions of most fraud samples and their non-fraud counterparts are the same. Within each pair of clusters, either the fraud samples cluster around their non-fraud counterparts, or the non-fraud samples cluster around their fraud counterparts. If such a spatial relationship among fraud and non-fraud samples does exist, the associated classification rule can be set up to identify the fraud samples based on the correspondence of the fraud samples and their non-fraud counterparts and vice versa. Moreover, the proposed dual approach is data-driven. That is, the corresponding system modeling is performed via directly using the sampled data. Thus, different sampled data input to the proposed DSS architecture may result in distinctive DSSs. To practically utilize such a spatial relationship for identifying fraud cases and examine the applicability of the proposed

DSS architecture based on the dual approach, this study sets up the fraudulent financial reporting (FFR) experiment, a sub-field of FFD.

Specifically, the proposed DSS architecture contains four phases. In the training phase, the sampling and variable selection are done first, and then it adopts the hierarchal-topology mapping advantage of the GHSOM to build up two GHSOMs (named non-fraud tree, NFT, and fraud tree, FT) from two classes of training samples collected from the financial statements.

In the modeling phase, following the majority principle, the corresponsive FT leaf node for each NFT leaf node are identified using all (fraud and non-fraud) training samples. Then, each training sample is classified into these two GHSOMs to develop the discrimination boundaries according to the candidate classification rules. The candidate classification rules in this study involve a non-fraud-central rule and a fraud-central rule, which are tuned via inputting the clustered training samples to determine the optimal discrimination boundary within each leaf node of the FT and NFT. For the candidate classification rules, a decision maker can set up his/her preference for the weights of classification error (type I and type II error) that makes the developed classification rule more acceptable and domain specific. The dominant classification rule with the best classification performance is applied in the analyzing phase. Besides, this study involves a feature extraction mechanism with two modules, feature-extracting module and pattern-extracting module, in the modeling phase that focus on discovering the common features and patterns in each FT leaf node. For the features regarding the input variables, the principal component analysis (PCA) is applied to provide the associated principal components. For the patterns such as the FFR fraud categories, the corresponding verdict contents of the fraud samples are investigated to determine the associated FFR fraud categories.

In the analyzing phase, each investigated sample is classified into the winning leaf nodes of FT and NFT, and applies the dominant classification rule to determine whether this sample is fraud or not. In the decision support phase, for an investigated sample, the result of the analyzing phase is used to help the decision makers speculate its FFR potentiality and. If it is identified as fraud, the associated potential FFR behaviors will be retrieved. The released information of the implemented DSS architecture based on the proposed dual approach can help decision makers better identify FFR and interpret the distinctive FFR behaviors among the clustered groups,

comprehend the difference between fraud and non-fraud samples, and finally facilitate the real-world decision making.

In sum, the implementation of the DSS architecture based on the proposed dual approach can be leveraged both to justify the spatial hypothesis, and when the spatial hypothesis holds, to disclose the information that better supports FFD. The proposed DSS architecture based on the dual approach is expected to be potentially applicable to other similar scenarios, and is able to be implemented as a DSS that helps detect suspicious samples and at the same time provide their possible fraud categories beforehand.

There are four objectives of this study:

(1) Develop a DSS architecture based on the proposed dual approach that (a) adopt the GHSOM to separately cluster fraud training samples and non-fraud training samples; (b) set up the discriminant boundaries for each pair of leaf nodes following the candidate classification rules based on the proposed spatial hypotheses; (c) use the determined classification rule to classify unknown samples; (d) observe whether spatial hypothesis holds and (d) illustrate the embedded information from the evaluation results including the extracted features and fraud patterns from the FT leaf nodes.

(2) Justify whether the implemented dual approach is capable of helping distinguish fraud and non-fraud samples.

(3) Compare the outcomes of our classification with other supervised or unsupervised learning methods.

(4) Provide implications regarding FFD decision support and research implications.

The rest of this dissertation is organized as follows. Chapter Two presents the literature reviews of the DSS, clustering methods, the GHSOM, PCA and FFR. Chapter Three explains the design of the proposed dual approach in details. Chapter Four demonstrates the experimental results. Chapter Five provides the comparison against other methods and the discussion of experiment of results. The implications are shown in Chapter Six. Chapter Seven gives the final conclusion and future works.

2. Literature review

In this section, we briefly review the DSS, clustering methods, GHSOM, PCA and FFR as the background knowledge including the applications of the GHSOM and the FFR detection issue.

2.1 DSS

Basically, the efforts in supporting the whole decision-making process focused in the development of computer information systems providing the support needed. The concept of the DSS was introduced, from a theoretical point of view, in the last 1960s. Klein and Methlie (1995) define a DDS as a computer information system that provides information in a specific problem domain using analytical decision models as well as techniques and access to database, in order to support a decision maker in taking decision effectively in complex and ill-structured problems. The contribution of DSS technology can be summarized as follows (Turban, 1993):

- The DSSs provide the necessary means for dealing with semi-structured and unstructured problems of high complexity, such as many problems from the field of financial management.
- The support provided by the DSSs may respond to the needs and the cognitive style of different decision makers, combining the preferences and the judgment of the every individual decision maker with the information derived by analytical decision models.
- The time and the cost of the whole decision process are significantly reduced.
- The support that is provided by the DSSs responds to the needs of various managerial levels, ranging from top managers and executives down to staff managers.

The DSSs help the decision maker to gain experience in data collection, as well as in the implementation of several scientific decision models, and they also incorporate the preferences and decision policy of the decision maker in the decision-making process. (Zopounidis et al., 1997)

There are various researches that have developed the DSSs for many application areas, for example, HR planning and decisions (Mohanty and Deshmukh, 1997),

financial management (Matsatsinis et al., 1997), marketing (Li, 2000), etc. Pinson (1992) developed the CREDEX system, which demonstrated the feasibility of a multi-expert approach driven by a meta-model in the assessment of credit risk. The system, using quantitative (economic and financial) and qualitative (social) data concerning the examined company and its business sector, as well as the bank's lending policy, provides a diagnosis of the company's function (commercial, financial, managerial and industrial) in terms of weaknesses and strengths. Zopounidis et al. (1997) developed a knowledge-based decision support system for financial management that integrates the DSS technologies to tackle past and current frequently occurring problems. Wen et al. (2005) proposed a decision support system based on an integrated knowledge base for acquisitions and mergers. It not only provided information concerning merger processes, major problems likely to occur in merger situations, and regulations practically or procedurally, but also gave rational suggestions in compliance with the appropriate regulations. It also suggested to the user how to deal with an uncertain growth rate and current evaluations. Wen et al. (2008) presented a mobile knowledge management decision support system using multi-agent technology for automatically providing efficient solutions for decision making and managing an electronic business. Nguyen et al. (2008) proposed an early warning system (EWS) that identifies potential bank failures or high-risk banks through the traits of financial distress which is able to identify the inherent traits of financial distress based on financial covariates (features) derived from publicly available financial statements.

In sum, many methodologies have been used and embedded with the existing framework of the DSS that could considerably increase the effectiveness of the provided decision support.

2.2 Clustering methods and the GHSOM

2.2.1 Clustering methods and the SOM

Clustering is an unsupervised classification of patterns into groups based on similarity. The main goal of clustering is to partition data patterns into several homogeneous groups that minimizes within-group variation and maximizes between

group variations. Each group is represented by the centroid of the patterns that belongs to the group. There are many important applications of clustering such as image segmentation (Jain et al., 1999), object recognition, information retrieval (Rasmussen, 1992), and so on. Clustering is the process of grouping the similarity data together such that data is high similarity within cluster but are dissimilarity between clusters. Clustering is the basis of many areas including data mining, statistical, biology, machine learning, etc. Clustering methods are used for data exploration and to provide class prototypes for use in the supervised classifiers. Among many clustering tools, the SOM is an unsupervised learning ANN and it appears to be an effective method for feature extraction and classification. Therefore, this study gives the following introduction and some literature reviews.

The Self-Organizing Map (SOM) is developed by Kohonen (1982), also known as the Kohonen Maps. It has demonstrated its efficiency in real domains, including clustering, the recognition of patterns, the reduction of dimensions, and the extraction of features. It maps high-dimensional input data onto a low dimensional space while preserving the topological relationships between the input data. SOM is made up two neural layers. The input layer has as many neurons as it has variables, and its function is merely to capture the information. Let m be the number of neurons in the input layer; and let $n_x * n_y$ the number of neurons in the output layer which are arranged in a rectangular pattern with x rows and y columns, which is called the map. Each neuron in the input layer is connected to each neuron in the output layer. Thus, each neuron in the output layer connections to the input layer. Each one of these connections has a synaptic weight associated with it. Let w_{ij} the weight associated with the connection between input neuron i and output neuron j. Figure 1 gives a visual representation of this neural arrangement.

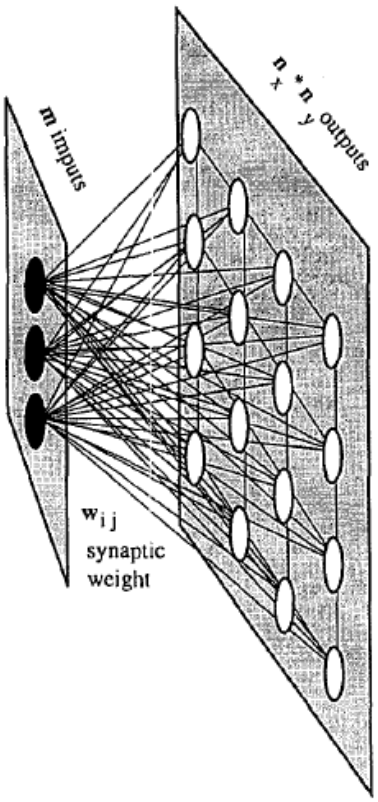


Figure 1. The self organizing map structure

Note: This SOM with m neurons in the input layer and n_x * n_y neurons in the output layer. Each neuron in the output layer has m connections w_{ij} (synaptic weights) to the input layer (Carlos, 1996).

SOM tries to project the multidimensional input space, which in our case could be financial information, into the output space in such a way that the input patterns whose variables present similar values appear close to one another on the map which is created. Each neuron learns to recognize a specific type of input pattern. Neurons which are close on the map will recognize similar input patterns whose images therefore, will appear close to one another on the created map. In this way, the essential topology of the input space is preserved in the output space. In order to achieve this, the SOM uses a competitive algorithm known as "winner takes all".

Initially, the w_{ij} are given random values. These values will be corrected as the algorithm progress. Training proceeds by presenting the input layer with financial ratios, one sample at a time. Let r_{ik} be the value of ratio i for firm k. This ratio will be read by neuron i. The algorithm takes each neuron in the output layer at a time and computes the Euclidean distance as the similarity measure.

$$d(j,k) = \sqrt{\sum_i (r_{ik} - w_{ijk})^2} \quad (1)$$

The output neuron for which $d(j, k)$ (defined in Equation (1)) is the smallest, and is the "winner neuron". Let such neuron be k^*. The algorithm now proceeds to change the synaptic weights w_{ij} in such a way that the distance $d(j, k^*)$ is reduced. A correction takes place, which depends on the number of iterations already performed and on the absolute value of the difference between r_{ij} and w_{ijk}. But other synaptic weights are also adjusted in function to how near they are to the winning neuron k^* and the number of iterations that have already taken place.

The procedure is repeated until complete training stops. Once the training is complete, the weights are fixed and the network is ready to be used. When a new pattern is presented, each neuron computes in parallel the distance between the input vector and the weight vector that it stores, and a competition starts that is won by the neuron whose weights are more similar to the input vector. Alternatively, we can consider the activity of the neurons on the map (inverse to the distance) as the output. The region where the maximum activity takes place indicates the class that the present input vector belongs to. If a new pattern is presented to the input layer and no neuron is stimulated by its presence the activity will be minimal, and this means that the pattern is not recognized. (Kohonen, 1989).

Thousands of the SOM applications are found among various disciplines (Serran, 1996; Richardson et al., 2003; Risien et al., 2004; Liu et al., 2006). It is widely used in application to the analysis of financial information (Serran, 1996). Eklund (2002) indicated that the SOM can be a feasible tool for classification of large amounts of financial data. The SOM has established its position as a widely applied tool in data-analysis and visualization of high-dimensional data. Within other statistical methods the SOM has no close counterpart, and thus it provides a complementary view to the data. The SOM is a widely used method in classification or clustering problem, because it provides some notable advantages over the alternatives (Khan et al., 2009).

There are various studies that used the SOM for a given clustering problem. Mangiameli, Chen, and West (1996) compared the performance of the SOM and seven hierarchical clustering methods for 252 data sets with various levels of imperfections that include data dispersion, outliers, irrelevant variables and non-uniform cluster densities. In conclusion, they demonstrated that the SOM is superior to the hierarchical

clustering methods. Granzow et al. (2001) investigated five clustering techniques: K-means, SOM, growing cell structure networks, fuzzy C-means (FCM) algorithm and fuzzy SOM. At the end of the analysis, they concluded that fuzzy SOM approach is the most suitable method in partitioning the data set. Shin and Sohn (2004) used K-means, SOM and FCM in order to segment stock trading customers and inferred that FCM cluster analysis is the most robust approach for segmentation of customers. Martín-Guerrero et al. (2006) compared the performance of K-means, FCM, and a set of hierarchical algorithms, Gaussian mixtures trained by the expectation–maximization algorithm, and the SOM in order to determine the most suitable algorithm in classification of artificial data sets produced for web portals. Finally, they concluded that the SOM outperforms the other clustering methods. Budayan et al. (2009) presented the strategic group analysis of Turkish contractors to compare the performances of traditional cluster analysis techniques, SOM and FCM for strategic grouping. It is concluded that the SOM and FCM can reveal the typology of the strategic groups better than traditional cluster analysis and they are more likely to provide useful information about the real strategic group structure.

The difference findings of these studies can be explained by the argument that the suitability of clustering methods to a given problem changes with the structure of the data set and purpose of the study. It is concluded that the aim of a study using clustering method is not to find out the best clustering method for all data sets and fields of application, but instead it is to demonstrate superior features of different clustering techniques for a particular problem domain, for example the FFD.

2.2.2 GHSOM

The SOM has shown to be a stable neural network model of high-dimensional data analysis. However, its capability is limited by some limitations when using the SOM. The first drawback is its static network architecture. The number and arrangement of nodes has to be pre-defined even without a priori knowledge of the data. Second, the SOM model has limited capabilities for the representation of hierarchical relations of the data. To overcome the inherent deficiencies of the SOM, Dittenbach, Merkl, and Rauber (2000) developed GHSOM to provide a SOM hierarchy automatically.

As shown in Figure 2, the GHSOM contains a number of SOMs in each layer. The size of these SOMs and the depth of the hierarchy are determined during its learning process according to the requirements of the input data.

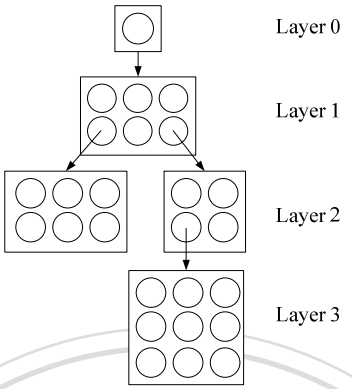


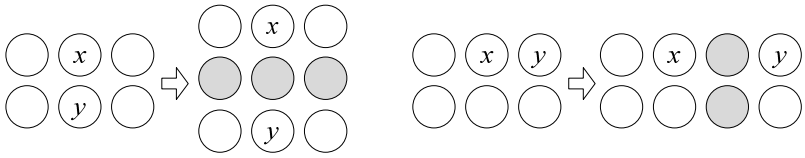
Figure 2. The GHSOM structure.

The training process of the GHSOM consists of the following four phases (Dittenbach et al., 2000):

- Initialize the layer 0: The layer 0 includes single node whose weight vector is initialized as the expected value of all input data. Then, the mean quantization error of layer 0 (MQE_0) is calculated. The MQE of a node denotes the mean quantization error that sums up the deviation between the weight vector of the node and every input data mapped to the node.
- Train each individual SOM: Within the training process of an individual SOM, the input data is imported one by one. The distances between the imported input data and the weight vector of all nodes are calculated. The node with the shortest distance is selected as the winner. Under the competitive learning principle, only the winner and its neighborhood nodes are qualified to adjust their weight vectors. Repeat the competition and the training until the learning rate decreases to a certain value.
- Grow horizontally each individual SOM: Each individual SOM will grow until the mean value of the MQEs for all of the nodes on the SOM (MQE_m) is smaller than the MQE of the parent node (MQE_p) multiplied by τ_1 as stated in Equation (2). If the stopping criterion is not satisfied, find the error node that

owns the largest MQE and then, as shown in Figure 3, insert one row or one column of new nodes between the error node and its dissimilar neighbor.

$$MQE_m < \tau_1 \times MQE_p \quad (2)$$



(a) insert a row (b) insert a column

Figure 3. Horizontal growth of individual SOM.

Note: The notation *x* indicates the error node and *y* indicates the *x*'s dissimilar neighbor.

- Expand or terminate the hierarchical structure: After the horizontal growth phase of individual SOM, each MQE_i is compared with the value of MQE_0 multiplied by τ_2. The node with an MQE_i greater than $\tau_2 \times MQE_0$ will develop a next layer of SOM. In this way, the hierarchy grows until all of the leaf nodes satisfy the stopping criterion stated in Equation (3). The leaf nodes means the node does not own a next layer of SOM.

$$MQE_i < \tau_2 \times MQE_0 \quad (3)$$

Several researches have applied the GHSOM to deal with text mining, image recognition and web mining problem. For example, Schweighofer et al. (2001) have show the feasibility of using the GHSOM and LabelSOM techniques in legal research by tests with text corpora in European case law. Shih et al. (2008) used the GHSOM algorithm to present a content-based and easy-to-use map hierarchy for Chinese legal documents in the securities and futures markets in the Chinese language. Antonio et al. (2008) used the GHSOM to analyze a citizen web portal, and provided a new visualization of the patterns in the hierarchical structure. The results have shown that the GHSOM is a powerful and versatile tool to extract relevant and straightforward knowledge from the vast amount of information involved in a real citizen web portal. Lu and Wang (2010) applied the GHSOM with support vector regression model to

product demand forecasting. The experimental results showed that the GHSOM can be used to combine with other machine learning or data mining techniques in order to improve the performance and obtain inspirable results.

Not many studies have applied the GHSOM in the purpose of forecasting until recent years. For instance, the two-stage architecture is employed by Hsu et al. (2009) which applied GHSOM and SVM to better predict financial indices. They suggested that the two-stage architecture can have smaller deviations between predicted and actual values than the single SVM model. Lu and Wang (2010) applied the GHSOM with support vector regression model to product demand forecasting. Guo et al. (2011) applied the GHSOM in case base reasoning system in design domain and found that new case is guided into corresponding sub-case base through the GHSOM, which makes the case retrieval more efficient and accurate. In sum, the GHSOM has been used in combining with other machine learning or data mining techniques to improve the model performance, and to provide valuable information for decision aid.

2.3 PCA

Feature extraction is an essential pre-processing step to pattern recognition and machine learning problems. It is often decomposed into feature construction and feature selection. Feature selection approaches try to find a subset of the original variables, which are generally performed before or after model training. In some cases, data analysis such as regression or classification can be done in the reduced space more accurately than in the original space. Feature selection can be done by using different methods, such as the PCA, Factor Analysis (FA), stepwise regression, and discriminant analysis (Tsai, 2009). In terms of the usage of dependent variable, these methods could be divided into supervised and unsupervised categories. Supervised feature selection techniques usually relate to the discriminant analysis technique (Fukunaga, 1990) which uses the within and between-class scatter matrices. Unsupervised linear feature selection techniques more or less all rely on the PCA (Pearson, 1901), which rotates the original feature space and projects the feature vectors onto a limited amount of axes (Turk and Pentland, 1991; Oja, 1992).

The PCA was invented by Pearson (1901). The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of uncorrelated principal components (PCs), which are ordered so that the first few retain most of the variation present in all of the original variables (Jolliffe, 2002).

The PCA can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute. The results of a PCA are usually discussed in terms of component scores (the transformed variable values corresponding to a particular case in the data) and loadings (the variance each original variable would have if the data are projected onto a given PCA axis) (Shaw, 2003).

The PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on (Jolliffe, 2002).

Define a data matrix, X^T, with zero empirical mean (the empirical mean of the distribution has been subtracted from the data set), where each of the n rows represents a different repetition of the experiment, and each of the m columns gives a particular kind of datum (say, the results from a particular probe). (Note that what we are calling X^T is often alternatively denoted as X itself.) The PCA transformation is then given by below Equation (4):

$$Y^T = X^T W = V\Sigma^T \tag{4}$$

where the matrices W, Σ, and V are given by a singular value decomposition (SVD) of X as $W \Sigma V^T$. (V is not uniquely defined in the usual case when $m < n-1$, but Y will usually still be uniquely defined.) Σ is an m-by-n diagonal matrix with nonnegative real numbers on the diagonal. Since W (by definition of the SVD of a real matrix) is an orthogonal matrix, each row of Y^T is simply a rotation of the corresponding row of X^T. The first column of Y^T is made up of the "scores" of the cases with respect to the principal component, and the next column has the scores with respect to the second principal component. If we want a reduced-dimensionality representation, we can

project X down into the reduced space defined by only the first L singular vectors, W_L defined in Equation (5):

$$Y = W_L{}^T X = \Sigma_L V_L{}^T \tag{5}$$

The matrix W of singular vectors of X is equivalently the matrix W of eigenvectors of the matrix of observed covariance $C = X X^T$ defined in Equation (6),

$$XX^T = W\Sigma\Sigma^T W^T \tag{6}$$

Given a set of points in Euclidean space, the first principal component corresponds to a line that passes through the multidimensional mean and minimizes the sum of squares of the distances of the points from the line. The second principal component corresponds to the same concept after all correlation with the first principal component has been subtracted out from the points. The singular values (in Σ) are the square roots of the eigenvalues of the matrix XX^T. Each eigenvalue is proportional to the portion of the "variance" (more correctly of the sum of the squared distances of the points from their multidimensional mean) that is correlated with each eigenvector. The sum of all the eigenvalues is equal to the sum of the squared distances of the points from their multidimensional mean. The PCA essentially rotates the set of points around their mean in order to align with the principal components. This moves as much of the variance as possible (using an orthogonal transformation) into the first few dimensions. The values in the remaining dimensions, therefore, tend to be small and may be dropped with minimal loss of information. The PCA is often used in this manner for dimensionality reduction. (Jolliffe, 2002)

The result of PCA is a linear transformation that transforms the data to a new coordinate system such that the new set of variables, also called the principal components. This linear function of the original variables are uncorrelated and the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. The main steps of the PCA are summarized in Figure 4.

- Calculate the covariance matrix or correlation matrix, C
- Compute the matrix V of eigenvectors which diagonalizes the covariance matrix C, where D is the diagonal matrix of eigenvalues of C. Matrix V, also of dimension M × M, contains M column vectors.

$$D = V^{-1}CV$$

$$Y^T = X^T W = V\Sigma^T$$

- Determination of number of significant components (L) based on statistical tests, variances limitation, or factor loadings
- Reproduction of Y using a reduced space defined by only the first L singular vectors, W_L

$$Y = W_L{}^T X = \Sigma_L \, V_L{}^T$$

Figure 4. The main steps of the PCA.

In short, the PCA is achieved by transforming to a new set of variables, as the principal components, which are uncorrelated and ordered so that the first few retain most of the variation present in the entire original variables (Jolliffe, 1986). By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables. Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped. (Ringnér, 2008)

Many studies have used the PCA for feature selection or dimensional reduction in financial studies. For example, Canbas et al. (2005) used the PCA to construct an integrated early warning system (IEWS) that can be used in bank examination and supervision process. In IEWS, the PCA helps explore and understand the underlying features of the financial ratios. By applying the PCA to the financial data, the important financial factors (i.e. capital adequacy, income-expenditure structure and liquidity), which can significantly explain the changes in financial conditions of the banks, were explicitly explored. Min and Lee (2005) reduced the number of multi-dimensional financial ratios to two factors through the PCA and calculate factor scores as the model training information. The result showed that the PCA contributes the graphic analysis step of support vector machines (SVMs) model with better explanatory power and stability to the bankruptcy prediction problem. Humpherys et al. (2010) applied the PCA with Varimax rotation and reliability statistics in their

proposed fraudulent financial detection model. Guided by theoretical insight and exploratory factor analysis, their 24-variable model of deception was reduced to a 10-variable model to achieve greater parsimony and interpretability.

Compare the PCA with FA, the PCA is preferred in this study because it is used to discover the empirical summary of the data set (Tabachnick and Fidell, 2001). In addition, the PCA considers the total variance accounting for all the common and unique (specific plus error) variance in a set of variables while FA considers only the common variance.

In the problem domain of FFD, the quantitative data are easier to present the financial conditions of the enterprise and an individual. This study tries to apply an analysis tool on quantitative clustered data to help to explore the represented variable sets and then give them a meaningful description. If the amount of sample is not much, the relationship between the input variables and output variable can be seen as linear; besides, we hope to find a composite of variables to provide more delicate group features. For this purpose, the PCA is more suitable and it has been widely used as a feature selection tool. Hence, this study will apply the PCA for feature extraction in our proposed dual approach in order to help get theoretical groups of input variables within each clustered group. That is, the PCA is used to provide expandability for each subgroup with clear endogenous variable insights; furthermore, these features can inspire the decision making process of the fraud detection, and can be enriched by other exogenous information related to fraud behaviors.

2.4 FFR

Fraudulent financial reporting (FFR), also known as financial statement fraud or management fraud, is a type of financial fraud that adversely affects stakeholders through misleading financial reports (Elliot and Willingham, 1980). FFR involves the intentional misstatement or omission of material information from an organization's financial reports (Beasley et al., 1999). FFR, although with the lowest frequency, casts a severe financial impact (Association of Certified Fraud Examiners, ACFE 2008). FFR can lead not only to significant risks for stockholders and creditors, but also financial crises for the capital market. According to the ACFE (2008), financial

misstatements are the most costly form of occupational fraud, with median losses of $2 million per scheme. FFR, or financial statement fraud, is known as “cooking the books” that often has severe economic consequences and makes front page headlines (Beasley et al., 1999). While ACFE (1998) reported that fraud has become more prevalent and costly, the detection of fraud has been badly lagging. The KPMG (1998) survey found that over one third of fraud cases were discovered by accident and that only 4 percent of cases were detected by independent auditor. When the auditor makes inquiries about fraud-related transactions, he or she is likely to be deceived with false or incomplete information (Weisenborn and Norris, 1997). Though the ability to identify fraudulent behavior is desirable, humans are only slightly better than chance at detecting deception (Bond and DePaulo, 2006) or identifying fraudulent behaviors beforehand. Therefore, there is an imperative need for decision aids of identifying FFR. More reliable methods are needed to assist auditors and enforcement officers in maintaining trust and integrity in publicly owned corporation.

Most prior FFR-related research focused on the nature or the prediction of FFR. The nature-related FFR research often uses the case study approach and provides a descriptive analysis of the characteristics of FFR and techniques commonly used. For example, the Committee of Sponsoring Organizations (COSO) and the Association of Certified Fraud Examiners (ACFE) regularly publish their own analysis on fraudulent financial reporting of U.S. companies. Based on the FFR samples, COSO examines and summarizes certain key company and management characteristics. ACFE analyzes the nature of occupational fraud schemes and provides suggestions to create adequate internal control mechanisms. As shown in Table 1, nature-related FFR research often uses case study, statistic or data mining approach to archival data and identifies significant variables that help predict the occurrence of fraudulent financial reporting. Other nature-related FFR studies focus on the audit assessment and planning (Bell and Carcello, 2000; Newman et al., 2001; Carcello and Nagy, 2004; Gillett and Uddin, 2005).

Table 1. Research methodology and findings in nature-related FFR studies.

Research	Methodology	Findings
Beasley et al. (1999)	• Case study • Descriptive statistics	• Nature of companies involved – Companies committing financial statement fraud were relatively small. – Companies committing the fraud were inclined to experience net losses or close to break-even positions in periods before the fraud. • Nature of the control environment – Top senior executives were frequently involved. – Most audit committees only met about once a year or the company had no audit committee. • Nature of the frauds – Cumulative amounts of fraud were relatively large in light of the relatively small sizes of the companies involved. – Most frauds were not isolated to a single fiscal period. – Typical financial statement fraud techniques involved the overstatement of revenues and assets. • Consequences for the company and individuals involved – Severe consequences awaited companies committing fraud. – Consequences associated with financial statement fraud were severe for individuals allegedly involved.

ACFE (2008)	• Case study • Descriptive statistics	• Occupational fraud schemes tend to be extremely costly. The median loss was $175,000. More than one-quarter of the frauds involved losses of at least $1 million. • Occupational fraud schemes frequently continue for years, two years in typical, before they are detected. • There are 11 distinct categories of occupational fraud. Financial statement fraud was the most costly category with a median loss of $2 million for the cases examined. • The industries most commonly victimized by fraud in our study were banking and financial services (15% of cases), government (12%) and healthcare (8%). • Fraud perpetrators often display behavioral traits that serve as indicators of possible illegal behavior. In financial statement fraud cases, which tend to be the most costly, excessive organizational pressure to perform was a particularly strong warning sign.

Another type of FFR research often uses the empirical approach to archival data and identifies significant variables that help predict the occurrence of FFR. This line of research also inputs these significant variables into fraud prediction models. Such research emphasizes the predictability of the model being used. For example, logistic regression and neural network techniques are used in this line of research (e.g., Persons, 1995; Fanning and Cogger, 1998; Bell and Carcello, 2000; Virdhagriswaran, 2006; Kirkos et al., 2007). The matched-sample design is typical for traditional FFR empirical studies. That is, a set of samples with fraudulent financial statements confirmed by the Department of Justice is matched with a set of samples without any allegations of fraudulent reporting.

Table 2 summarizes the research methodology and findings of the FFR empirical studies most relevant to our study. The research methodology has shown a trend with an emphasis on the classification mechanization which is used as the decision support information for future risk identification (Basens et al., 2003).

Table 2. Research methodology and findings in FFR empirical studies.

Author	Methodology	Variable	Sample	Findings
Dechow et al. (1996)	Logistic regression	• 21 variables – Financial ratios – Other indicators: corporate governance ratios.	Matched-pairs design: 92 firms subject to enforcement actions by the SEC	• To attract external financing at low cost was found an important motivation for earnings manipulation • Firms manipulating earnings are more likely to have: - insiders dominated boards - Chief Executive Officer simultaneously serves as Chairman of the Board
Persons (1995)	Stepwise logistic model	• 9 financial ratios • Z-score	Matched-pairs design	The study found four significant indicators: financial leverage, capital turnover, asset composition and firm size
Fanning and Cogger (1998)	Self-organizing artificial neural network	• 62 variables • Financial ratios • Other indicators: corporate governance, capital structure etc.	Matched-pairs design: 102 fraud samples and 102 non-fraud samples	• Neural network is more effective • Financial ratios such as debt to equity, ratios of accounts receivable to sales, trend variables are significant indicators
Bell and Carcello (2000)	Logistic regression	46 fraud risk factors	77 fraud samples and 305 non-fraud samples	Logistic regression model outperformed auditors for fraud samples, but were equally performed for non-fraud samples.
Kirkos et al. (2007)	• Decision tree • Back-propagation neural network • Bayesian belief	• 27 financial ratios • Z-score	Matched-pairs design: 38 fraud samples and 38 non-fraud	• Training dataset: neural network is the most accurate • Validation dataset: Bayesian belief network is

	network		samples	the most accurate
Hoogs et al. (2007)	Genetic Algorithm	• 38 financial ratios • 9 qualitative indicators	51 fraud samples vs. 51 non-fraud samples	Integrated pattern had a wider coverage for suspected fraud companies while it remained lower false classification rate for non-fraud ones

Source: (Hsu, 2008; Huang et al., 2011).

As shown in Table 2, Persons (1995) used Stepwise logistic model to found significant indicators relate to FFR. Dechow et al. (1996) used Logistic regression in FFR detection. Bell and Carcello (2000) developed a Logistic regression model useful in predicting the existence of fraudulent financial reporting, and found that the proposed model outperformed auditors for fraud samples, but were equally performed for non-fraud samples.

Green and Choi (1997) applied Back-propagation neural network to FFR detection. The model used five ratios and three accounts as input. The results showed that Back-propagation neural network had significant capabilities when used as a fraud detection tool. Fanning and Cogger (1998) proposed a generalized adaptive neural network algorithm, named AutoNet, to FFR detection. The input vector consisted of financial ratios and qualitative variables. They compared the performance of their model with linear and quadratic discriminant analysis, as well as logistic regression, and claimed that AutoNet is more effective at detecting fraud than standard statistical methods. Kirkos et al. (2007) compared Decision tree, Back-propagation neural network, and Bayesian belief network in FFR detection and found that Back-propagation neural network is the most accurate method in training dataset, Bayesian belief network is the most accurate method in validation dataset. Hoogs et al. (2007) applied Genetic Algorithm (GA) in FFR detection, and the performance of GA concluded that the integrated pattern had a wider coverage for suspected fraud companies while it remained lower false classification rate for non-fraud ones.

Humpherys et al. (2010) developed a linguistic methodology for detecting fraudulent financial statements. The results demonstrate that linguistic models of deception are potentially useful in discriminating deception and managerial fraud in

financial statements. Their findings provide critical knowledge about how deceivers craft fraudulent financial statements and expand the usefulness of deception models beyond a low-stakes, laboratory setting into a high-stakes, real-world environment where large fines and incarceration are the consequences of deception. In literature of financial fraud detection (FFD), Ngai et al. (2010) have done a complete classification framework and an academic review of literature which used data mining techniques for FFD. They showed that 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. Huang et al. (2011) used the GHSOM to help capital providers examine the integrity of financial statement. They applied the GHSOM to analysis financial data and demonstrate an alternative way to help capital providers such as lenders to evaluate the integrity of financial statements, a basis for further analysis to reach prudent decisions. Huang and Tsaih (2012) evolved the GHSOM into a prediction model for detecting the FFR. They proposed the initial concept of a dual approach for examining whether there is a certain spatial relationship among fraud and non-fraud samples, identifying the fraud counterpart of a non-fraud subgroup, and detecting fraud samples.

The relevant literatures show that the neural network families have been widely used in many financial applications, such as the FFR detection, credit ratings, economic forecasting, risk management, or other FFD related issues.

2.5 Summary

The GHSOM is an improved vision of SOM. It is often used as a clustering tool and has proved its availability to deal with classification and clustering problem to achieve the decision support purpose. As a clustering tool, the related GHSOM studies nowadays still provide limit information (or lack of the inherent knowledge) from the clustering results, which may increase decision makers' efforts to analyze such semi-structured results.

As a result, further analysis for the generated subgroups is needed. A particular design of the GHSOM for FFD is also needed since the learning nature of the GHSOM

is unsupervised. Recent researches which considered feature extraction or pattern recognition of the GHSOM are often applied to graphic data, sensor-collected data, or text content; however, few of studies have focus on financial data. Despite the GHSOM provides more delicate clustering results than the SOM, we find that no study has applied the GHSOM and integrated it into a DSS that helps identify fraud (e.g., FFR). Therefore, this study expects to utilize the advantage of the GHSOM to design a novel dual approach and apply it to detect FFR (a sub-field of FFD) that helps identify fraud cases and explore their imbedded features through the PCA and explore their potential FFR patterns through any qualitative method, and finally provides abundant detection results as the investigative report to facilitate the decision process of both identification and interpretation.

3. The proposed dual approach

The overall system architecture of the proposed dual approach is illustrated in Figure 5. The proposed dual approach consists of the following four phases: the training, the modeling, the analyzing, and the decision support. There are eleven major modules: sampling, variable-selecting, clustering, statistic-gathering, rule-forming, feature-extracting, pattern-extracting, classifying, analyzing, feature-retrieving, and decision-supporting.

The training phase consists of a series of three modules, which aims to sequentially sample the data, select the input variables, and set up two GHSOM trees based upon the dichotomous categories of training samples. The modeling phase consists of a series of four modules, which aims to calculate two statistical values (Avg and Std) from each leaf node of the obtained two GHSOM trees to form the optimal classification rule based on the training samples, and extract features using quantitative and qualitative method from the GHSOM tree which consists of fraud samples. In other word, the modeling phase mainly focuses on setting up the classification rule based on certain spatial relationship, which match each leaf node of FT to its counterpart leaf nodes in NFT and vice versa.

The analyzing phase consists of two modules, in which the set GHSOMs, the classification rule and the extracted features are used to identify fraud from the unknown investigated samples, and retrieve the associated features for decision aid. Each investigated sample will be classified into its belonging leaf node of GHSOMs. Then, the optimal classification rule will be used to classify the investigated sample. The decision support phase consists of two modules, which present the classification result and retrieve the fraud related features for decision aid. For an investigated sample, if the classification result is fraud, the regularity of its belonging leaf node of FT, that is the principal components and the potential fraud categories, are retrieved.

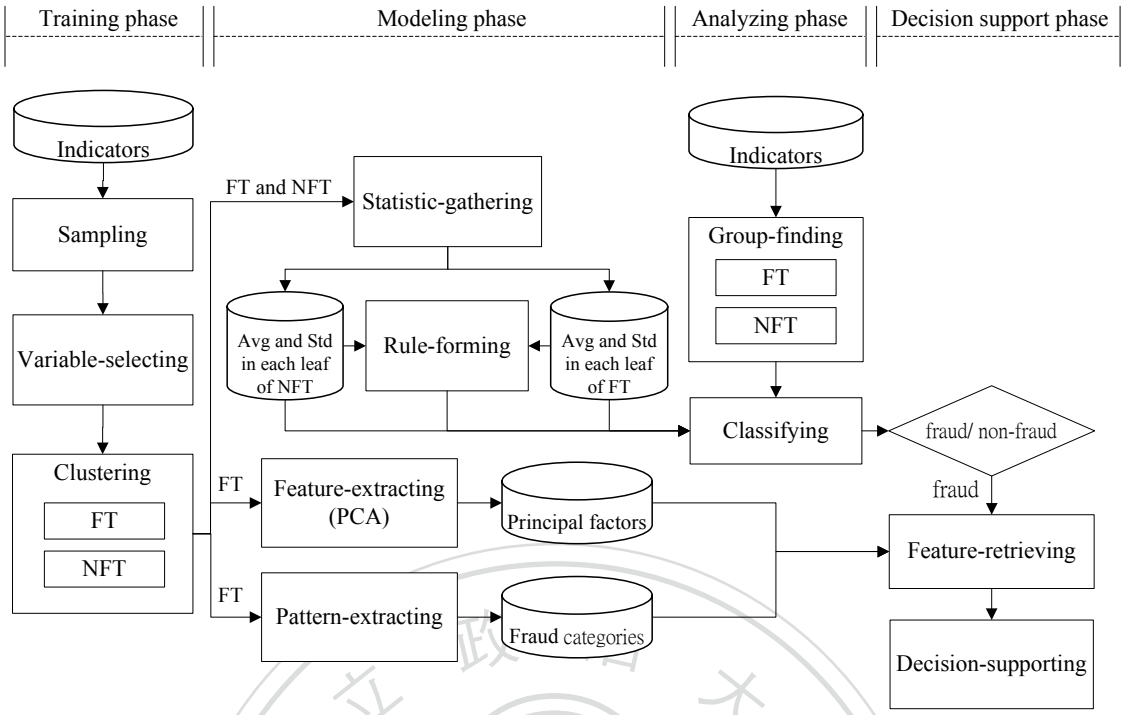


Figure 5. System architecture of the proposed dual approach.

Note: The clustering module generates two GHSOM trees. One is FT (use fraud samples), and the other one is NFT (use non-fraud samples). The group-finding module classifies the investigated sample into FT and NFT, respectively.

3.1 Training phase

Table 3 shows the training phase, in which the task of data pre-processing is done via step 1 and step 2. Step 2 can apply any variable selection tool such as the discriminant analysis, logistic model, and so forth.

Since fraud and non-fraud samples will be used to set up two GHSOM trees (named non-fraud tree, NFT, and fraud tree, FT) respectively, before processing step 3, the training samples are grouped as the fraud ones and the non-fraud ones. In step 3, the fraud samples are used to set up an acceptable GHSOM named FT. After identifying the FT, the values for (the GHSOM's) breadth parameter (τ_1) and depth parameter (τ_2) are determined and stored in step 4. Then, in step 5, the determined values of τ_1 and τ_2 and the non-fraud samples are used for setting up another GHSOM named NFT.

Table 3. The training phase.

step 1:	Sample and measure variable.
step 2:	Identify the significant variables that will be used as the input variables.
step 3:	Use the fraud samples to set up an acceptable GHSOM (denote FT).
step 4:	Based upon the accepted FT, determine the (GHSOM training) parameters breadth (τ_1) and depth (τ_2).
step 5:	Use the non-fraud samples and the determined parameters τ_1 and τ_2 to set up another GHSOM (denote NFT).

3.1.1 Sampling module

The sampling module processes sample collection and variable measurement. The sampling module is executed via the step 1 of Table 3. The definition of a fraud sample and a non-fraud sample are defined first. The sources, the sample period and the way of sampling are also decided in this step. The design of the sampling process is flexible depended on the application field.

The explanatory variables are selected based upon fraud related literatures. These measurements may proxy for several attributes of a sample. The next step will help to select variables significantly relate to fraud, which help downsize the number of input variables to make it more relevant to the collected sample base.

3.1.2 Variable-selecting module

The variable-selecting module is executed via the step 2 of Table 3. In the variable-selecting module, the collected explanatory variables and the fraud/non-fraud dichotomous dependent variable are put into the variable selection tool. Any variable selection tool such as discriminant analysis, logistic model, and so forth can be applied in this step. For example, in this study, the variable-selecting module applies discriminant analysis processing variable selection from the obtained samples in order to identify the significant variables that help detect fraud. Then, the variables with statistically significant effects will be chosen as the input variables for GHSOM training to obtain clustered groups.

3.1.3 Clustering module

The clustering module is executed via the step 3, step 4 and step 5 of Table 3. The significant variables derived from the variable-selecting module are used as the input variables for the GHSOM training to conduct clustering procedure. Two GHSOMs (named non-fraud tree, NFT, and fraud tree, FT) are respectively generated from two classes of training samples (fraud class, non-fraud class). For each GHSOM, a series of leaf nodes (i.e., groups) are generated. Furthermore, based upon the FT, we can get several clustered groups with inherent similarity for helping further feature extraction.

The competitive learning nature of GHSOM makes it work as a regularity detector that is supposed to discover statistically salient features of the sample population (Rumelhart and Zipser, 1985). In this module, the GHSOM will develop the topological representation which captures the most salient features of each cluster. Furthermore, through a set of small-sized leaf nodes, the GHSOM classifies the sample into more subgroups with hierarchical relations instead of a dichotomous result and therefore further delicate analyses are feasible.

3.2 Modeling phase

Table 4 presents the modeling phase, in which the classification rule is set up. Despite FT and NFT are resulted from fraud and non-fraud samples respectively, the spatial relationship hypothesis and the same setting of (τ_1 and τ_2) parameters may render true that each leaf node of NFT has one or more than one counterpart leaf nodes in FT and vice versa. Thus, one purpose of the modeling phase is to match each leaf node of FT to its counterpart leaf nodes in NFT and vice versa.

Table 4. The modeling phase.

step 1: For each leaf node of FT, i. calculate and store its Avg_x value that is the average of Euclidean distances between the weight vector and the grouped fraud training samples; ii. calculate and store its Std_x value that is the standard deviation of

Euclidean distances between the weight vector and the grouped fraud training samples.

step 2: For each leaf node of NFT,

i. calculate and store its Avg_y value that is the average of Euclidean distances between the weight vector and the grouped non-fraud training samples;

ii. calculate and store its Std_y value that is the standard deviation of Euclidean distances between the weight vector and the grouped non-fraud training samples.

step 3: For each training sample,

i. identify and store the winning leaf node of FT and the winning leaf node of NFT, respectively;

ii. store its *Avg* values of the winning leaf nodes of FT and NFT, respectively;

iii. store its *Std* values of the winning leaf nodes of FT and NFT, respectively.

iv. calculate and store its D_{ft}, the Euclidean distance between the training sample and the weight vector of the winning leaf node of FT;

v. calculate and store its D_{nft}, the Euclidean distance between the training sample and the weight vector of the winning leaf node of NFT.

step 4: Create the spatial correspondence tables regarding the matching from NFT to FT and from FT and NFT, respectively.

step 5: Use the fraud-central rule defined in Equation (3) and the optimization problem (4) to determine the parameter β_1^p that minimizes the corresponding sum of (type I and type II) classification errors.

step 6: Use the non-fraud-central rule defined in Equation (7) and the optimization problem (8) to determine the parameter β_2^p that minimizes the corresponding sum of (type I and type II) classification errors.

step 7: Pick up the dominant classification rule via comparing the classification errors obtained in step 5 and step 6.

step 8: For each leaf node of FT, apply PCA to select features through extracting factors (i.e., principle components).

step 9: For each leaf node of FT, analyze the common fraud features from exogenous information based on the associated domain categories.

3.2.1 Statistic-gathering module

The statistic-gathering module is executed via the step 1, step 2 and step 3 of Table 4. After NFT and FT are constructed, a non-fraud-central rule and a fraud-central rule are tuned respectively via inputting all samples to determine the adjustable discrimination boundary within each leaf node of the NFT and FT. The optimization renders rules for detecting fraud samples are adjustable and effective. The decision maker can objectively set his/her weightings of type I and type II errors. The rule associated with the tree that dominates another is adopted as the classification rule to classify whether samples are fraud or non-fraud.

In step 1, the Avg_x value (i.e., the average of Euclidean distances between the weight vector and the grouped fraud training samples) and the Std_x value (i.e., the standard deviation of Euclidean distances between the weight vector and the grouped fraud training samples) of each leaf node of FT are calculated and stored. Similarly, in step 2, the Avg_y value (i.e., the average of Euclidean distances between the weight vector and the grouped non-fraud training samples) and the Std_y value (i.e., the standard deviation of Euclidean distances between the weight vector and the grouped non-fraud training samples) of each leaf node of NFT are calculated and stored. Hereafter, we use #x to denote the x^{th} leaf node of FT and *y the y^{th} leaf node of NFT.

In step 3, we collect and store the following information regarding each training sample: the winning leaf node of FT, the winning leaf node of NFT, the corresponding Avg_x and Std_x values of the winning leaf node of FT, the corresponding Avg_y and Std_y values of the winning leaf node of NFT, the D_{ft} (i.e., the Euclidean distance between the training sample and the weight vector of the winning leaf node of FT), and D_{nft} (i.e., the Euclidean distance between the training sample and the weight vector of the winning leaf node of NFT). Following the GHSOM classification rule, we identify the winning leaf nodes of FT and NFT, respectively.

3.2.2 Rule-forming module

The rule-forming module is executed via the step 4 to step 7 of Table 4. In step 4, two spatial correspondence tables are created respectively based on the classification results of all (fraud and non-fraud) training samples. That is, from the NFT perspective, if the leaf node #*x* in FT hosts the majority of all training samples classified in the leaf node **y* in NFT, then we match the leaf node #*x* in FT to the leaf node **y* in NFT and claim that the leaf node #*x* in FT is the counterpart of the leaf node **y* in NFT. The leaf node matching of #*x* to **y* states the spatial relationship among the fraud and non-fraud samples classified in the leaf nodes of #*x* and **y*. That is, if any sample is classified into the leaf node **y* when using NFT, it is more likely to be classified into the leaf node #*x* when using FT. Similarly, from the FT perspective, if the leaf node **y* in NFT hosts the majority of all of training samples classified in the leaf node #*x* in FT, then we match the leaf node **y* in NFT to the leaf node #*x* in FT and claim that the leaf node **y* in NFT is the counterpart of the leaf node #*x* in FT.

The fraud-central rule defined in Equation (7), in which β_1^p is a parameter for a pair *p* of leaf nodes (#FT match to *NFT), states that some non-fraud samples cluster around a subset of fraud samples. That is, for the (fraud or non-fraud) sample *c* that is classified into the leaf node #*x* of FT, if D_{ft}^c is smaller than the value of $Avg_x^c + \beta_1^p \times Std_y^c$, the sample *c* will be classified as the fraud one; otherwise, the non-fraud one. Because the discrimination boundary (i.e., $Avg_x^c + \beta_1^p \times Std_y^c$) is data-dependent, the parameter β_1^p needs to be tuned to find the optimal discrimination boundary. Therefore, in step 5, we use the optimization problem (8) to determine the parameter β_1^p. In the optimization problem (8), the sets S_F and S_{NF} are given. For each *c*, the values of D_{ft}^c, Avg_x^c, and Std_y^c, are also given. In the objective function, there are coefficients w_1 (the weighting of type I error) and w_2 (the weighting of type II error) that are constants subjectively determined by the decision makers in terms of their preference of the classification performance. In general, there are three kinds of settings for (w_1, w_2) — (1, 1), (0.01, 1), (1, 0.01) regarding the minimizations focusing on the average sum of type I and type II errors, mainly the type II error, and mainly the type I error, respectively.

The fraud-central rule: If ($D_{ft}^c < Avg_x^c + \beta_1^p \times Std_y^c$), the sample is classified as the fraud one; otherwise, the non-fraud one. (7)

$$\min_{\beta_1} w_1 \times \sum_{c \in S_{NF}} (i^c + 1)^2 + w_2 \times \sum_{c \in S_F} (i^c - 1)^2 \tag{8}$$

$$s.t.\ i^c = \begin{cases} 1 & if \quad D_{ft}^c < Avg_x^c + \beta_1^p * Std_y^c \\ -1 & if \quad D_{ft}^c \geq Avg_x^c + \beta_1^p * Std_y^c \end{cases} \quad \text{for all } c \text{ in } S_F \text{ and } S_{NF}$$

From the definition of i^c, $(i^c)^2$ equals 1. Thus, the objective function can be refined as Equation (9) and, effectively, we only need to minimize

$w_1 \times \sum_{c \in S_{NF}} i^c$ - $w_2 \times \sum_{c \in S_F} i^c$ by varying β_1^p.

$$w_1 \times \sum_{c \in S_{NF}} (i^c + 1)^2 + w_2 \times \sum_{c \in S_F} (i^c - 1)^2 =$$

$$(|S_{NF}| + 1) \times w_1 + (|S_F| + 1) \times w_2 + 2w_1 \times \sum_{c \in S_{NF}} i^c - 2w_2 \times \sum_{c \in S_F} i^c \tag{9}$$

Since the values of D_{ft}^c, Avg_x^c, and Std_y^c remain unchanged as β_1^p varies, the following enumeration scheme can be used to determine the optimal values of β_1^p. Note that all Std_y^c are strictly positive. Thus we have:

$$i^c = \begin{cases} 1 & if \quad \dfrac{D_{ft}^c - Avg_x^c}{Std_y^c} < \beta_1^p \\ -1 & if \quad \dfrac{D_{ft}^c - Avg_x^c}{Std_y^c} \geq \beta_1^p \end{cases} = \begin{cases} 1 & if \quad e_{xy}^c < \beta_1^p \\ -1 & if \quad e_{xy}^c \geq \beta_1^p \end{cases} \tag{10}$$

where $e_{xy}^c \equiv \dfrac{D_{ft}^c - Avg_x^c}{Std_y^c}$ is a constant given c, x, and y. For all c in the set S_{NF}, arrange the values of e_{xy}^c in an increasing order. If the value of β_1^p is in the range less than the first e_{xy}^c, then $\sum_{c \in S_{NF}}$ = - $|S_{NF}|$. If the value of β_1^p is between the first and the second e_{xy}^c, then $\sum_{c \in S_{NF}} i^c = -|S_{NF}| + 2$. There are only $|S_{NF}|$ amount of e_{xy}^c values and, as β_1^p varies, we know the exact value of $w_1 \times \sum_{c \in S_{NF}} i^c$. Repeat the same process for the set S_F and there are only $|S_F|$ amount of e_{xy}^c values for us to check the value of $w_2 \times \sum_{c \in S_F} i^c$. By superimposing the two sequences of e_{xy}^c we can get the optimal ranges of β_1^p that minimizes the value of $w_1 \times \sum_{c \in S_{NF}} i^c$ - $w_2 \times \sum_{c \in S_F} i^c$.

The non-fraud-central rule defined in Equation (11), in which β_2^p is a parameter for a pair *p* of leaf nodes (*NFT match to #FT), states that some fraud samples cluster around a subset of non-fraud samples. That is, for the sample *c* that is classified into the leaf node **y* of NFT, if D_{nft}^c is smaller than the value of $Avg_y^c + \beta_2^p \times Std_x^c$, the sample *c* will be classified as the non-fraud one; otherwise, the fraud one. The parameter β_2^p also needs to be tuned to find the optimal discrimination boundary (i.e., $Avg_y^c + \beta_2^p \times Std_x^c$). Therefore, in step 6, we use the optimization problem stated in (8) to determine the parameter β_2^p through the minimization of the sum of (type I and type II) classification errors. In the optimization problem (12), the sets S_F and S_{NF} are given. For each *c*, the values of D_{nft}^c, Avg_y^c, and Std_x^c ,are also given. The constants w_1 and w_2 in the objective function are set as the same values as in optimization problem (8).

The non-fraud-central rule: If ($D_{nft}^c < Avg_y^c + \beta_2^p \times Std_x^c$), the sample is classified as the non-fraud one; otherwise, the fraud one. (11)

$$\min_{\beta_2} w_1 \times \sum_{c \in S_{NF}} (i^c - 1)^2 + w_2 \times \sum_{c \in S_F} (i^c + 1)^2 \quad (12)$$

$$s.t.\ i^c = \begin{cases} 1 & if \quad D_{nft}^c < Avg_y^c + \beta_2^p * Std_x^c \\ -1 & if \quad D_{nft}^c \geq Avg_y^c + \beta_2^p * Std_x^c \end{cases} \text{ for all } c \text{ in } S_F \text{ and } S_{NF}$$

The approach for solving the optimization problem (8) is also applied for solving the optimization problem (12) to get the optimal range of β_2^p that minimizes the value of $w_1 \times \sum_{c \in S_{NF}} (i^c - 1)^2 + w_2 \times \sum_{c \in S_F} (i^c + 1)^2$.

In step 7 of Table 4, the picked classification rule is the fraud-central rule if the sum of classification errors resulted in step 5 is smaller than the one resulted in step 6; otherwise, the non-fraud-central rule. The dominance of the non-fraud-central rule leads to an implication of the spatial relationship among fraud and non-fraud samples that most of fraud samples cluster around the non-fraud counterpart. The dominance of the fraud-central rule leads to an implication of the spatial relationship among fraud and non-fraud samples that most of non-fraud samples cluster around the fraud counterpart.

3.2.3 Feature-extracting module

The feature-extracting module is executed via the step 8 of Table 4. For each clustered group based upon fraud samples, the feature-extracting module applies PCA

to select features or to extract factors (i.e., principle components) that link to fraud related features from exogenous information. It further represents the inherent variable features to reveal each group's heterogeneity, and the purpose of feature selection is trying to exclude variables irrelevant to the modeling problem for a particular group. Here we use PCA to do feature selection by selecting a set of variables which best represent the composited features of an investigated leaf node of the GHSOM clustering result based upon fraud samples.

The main objective of the PCA is to determine the important dimensions (characters) which can explain the input variable features of the analyzed samples, and can explore underlying patterns of relationship between the input variables. The input variables are same as the GHSOM input variable. The fraud/ non-fraud dichotomous variable is set to the dependent variable. Only those factors that account for variances greater than 1 (eigenvalue >1) are included in the model. This criterion is also called K1 method proposed by Kaiser (1960) and is probably the one most widely used. According to this rule, only the factors that have eigenvalues greater than one are retained for interpretation. Factors with variance less than one are not better than a single ratio, since each ratio has a variance of 1.

The other objective of the PCA is to calculate factor scores for each of the sample according to the factors determined. Then, to enhance the interpretability of the factors, the varimax factor rotation method is used in PCA. This method minimizes the number of variables that have high loadings on a factor, and all factor loadings will be presented. Here, variables with large loadings for the same factors are grouped and small factor loadings are omitted. Estimated factor represents a specific characteristic of firms under consideration. (Canbas et al., 2005)

The outcomes of feature-extracting module are several representative variables as the 'variable pattern' for each clustered group. Hence, from comparing the similarity of each selected features provided by PCA, we can efficiently exploit one single group or compare different groups. Besides, after determining the basic financial factors from training samples, early warning model can be estimated according to these obtained factors, such as discriminant, logit, probit, and Neural Network.

3.2.4 Pattern-extracting module

The pattern-extracting module is executed via the step 9 of Table 4. The exogenous information of the fraud behaviors beyond the financial numbers is used in this module. Extracting the fraud categories of a certain investigated sample can help reveal more domain information. We can use any qualitative method to analyze the category of fraud from any available structural, semi-structural, or un-structural resource, such as news, reports, or other fraud-related content. First, the categories of fraud should be determined by the authentic reference. Then, for a leaf node of FT, using any qualitative way to classify the fraud categories of each samples belong to the leaf node. If the resource of fraud categories is structural, we only have to encode the class data as the other extracted feature.

3.3 Analyzing phase

The analyzing phase is shown in Table 5. For each investigated sample *s*, we first follow the GHSOM clustering rule to find the winning leaf nodes of FT and NFT, respectively. The indicators of the investigated samples are based upon the result of the variable-selecting module. Assume the winning leaf node of FT is the #*x* one and the winning leaf node of NFT is the **y* one. Then, we use the classification rule picked from the modeling phase to do the identification. That is, if the fraud-central rule is picked in the modeling phase, step 2 is processed via Equation (3) in the analyzing phase to classify the investigated sample *s*. If the non-fraud-central rule is picked in the modeling phase, step 3 is here processed via Equation (7) in the analyzing phase to classify the investigated sample *s*.

Table 5. The analyzing phase.

step 1:	For each investigated sample *s*, identify the winning leaf node #*x* of FT and the winning leaf node **y* of NFT, respectively.

step 2:	If the classification rule is the fraud-central rule, then
	i. Calculate D_{ft}, the Euclidean distance between the investigated sample *s* and the weight vector of the leaf node #*x* of FT.
	ii. Use the fraud-central rule with the determined β_1 value to classify the investigated sample *s*.
	If the classification rule is the non-fraud-central rule, then
	i. Calculate D_{nft}, the Euclidean distance between the investigated sample *s* and the weight vector of the leaf node **y* of NFT.
	ii. Use the non-fraud-central rule with the determined β_2 value to classify the investigated sample *s*.

3.3.1 Group-finding module

The group-finding module and the classifying module are implemented after setting up the classification rules based on the training samples. The group-finding module is executed via the step 1 of Table 5.

The group-finding module is processed to identify the belonging leaf node of FT and NFT for an investigated sample. Each investigated sample *s* is respectively imported to each leaf node of GHSOMs (FT and NFT) to classify it into the most similar leaf node. The input variables of the investigated samples are same as the clustering module, that are the chosen variables generated from the variable-selecting module. The distances between the imported input data and the weight vector of all leaf nodes of the GHSOM are calculated. The leaf node with the shortest distance is selected as the belonging leaf node.

3.3.2 Classifying module

The classifying module is executed via the step 2 and step 3 of Table 5. In the classifying module, the rule associated with the tree that dominates another in modeling phase is adopted as the classification rule to classify whether samples are fraud or non-fraud. The dominance of the non-fraud-central rule leads to an implication that

most of fraud samples cluster around the non-fraud counterpart, meanwhile the dominance of fraud-central rule leads to an implication that most of non-fraud samples cluster around the fraud counterpart.

3.4 Decision support phase

The decision support phase is shown in Table 6. This phase mainly provides the classification result and the fraud related features for any susceptive sample to the decision makers.

Following the SOM theories, common fraud categories and relevant variables extracted from the GHSOM clustering results are applicable to all samples clustered in the same FT leaf nodes. In the decision support phase, the associated features which are extracted from modeling phase will be retrieved by the feature-retrieving module and integrated by the decision-supporting module for further decision support. It not only classifies whether an investigated sample is fraud or not, but also tries to identify its potential committed fraud categories. The integrated information shell provide adequate resources to facilitate the decision making process.

Table 6. The decision support phase.

step 1:	Retrieve the associated fraud categories and principle components of the investigated sample *s*.
step 2:	Summarize the decision support results of the investigated sample *s*.

* The investigated sample *s* is classified as a fraud observation.

3.4.1 Feature-retrieving module

The feature-retrieving module is executed via the step 1 of Table 6. After obtaining the classification result for a particular unknown investigated sample, the associated features will be retrieved and integrated by the feature-retrieving module. The associated features of its belonging leaf node which include the principal factors extracted by PCA, and the potential fraud categories (techniques) extracted by any

qualitative method, are respectively retrieved from the outputs of the feature-extracting module and the pattern-extracting module.

Specifically, if an investigated sample is identified as a fraud one, its potential fraud categories and other associate features will be retrieved based on the GHSOM classification result of the group-finding module. On the contrary, if an investigated sample is not classified fraud, there is no need to retrieve the potential fraud categories. As a result, we can expect that the common fraud categories of a certain leaf node of FT represent the speculation for any investigated sample classified into this leaf node, and that contribute to the explanation of FFD result which is helpful for providing decision support.

3.4.2 Decision-supporting module

The decision-supporting module is executed via the step 2 of Table 6. It is developed to build up a decision support mediator which provides prediction summary retrieved from the result of classifying module and the associated features gathered by the feature-retrieving module. The classification result as well as the associated features of an investigated sample are summarized and provided for decision makers.

Such summary of a certain investigated sample requested by a decision maker will give him/ her completed information with clear background features and insights in terms of any potential fraud behavior derived from the modeling phase, that possess more traceable fraud knowledge than traditional fraud prediction models. For FFD detection purpose, this information can facilitate the decision support of fraud identification which reveals both fraud/non-fraud classification result and any potential fraudulent activity as a reference for further investigation.

In sum, a decision maker can get an evaluation result which consists of fraud/ non-fraud classification result, and the potential fraud categories of an investigated sample. If needed, they can view other samples which are classified into the same leaf node to get more background information. They can also view the whole GHSOM (e.g., FT) structure to understand the contrasted location between groups. These results provide decision makers an easy way of understanding the general picture of sample data. Connecting several features related to problem domain also helps decision makers get some insights quickly; besides, the reasonability of the detecting result can be

checked here to make sure if the following fraud prevention strategy is feasible. The concept of the training, modeling and analyzing phase of the dual approach is shown in the following Figure 6, in which the main process is depicted in each phase to help understand the ideas of each phase depicted in Table 3, Table 4, Table 5 and Table 6.

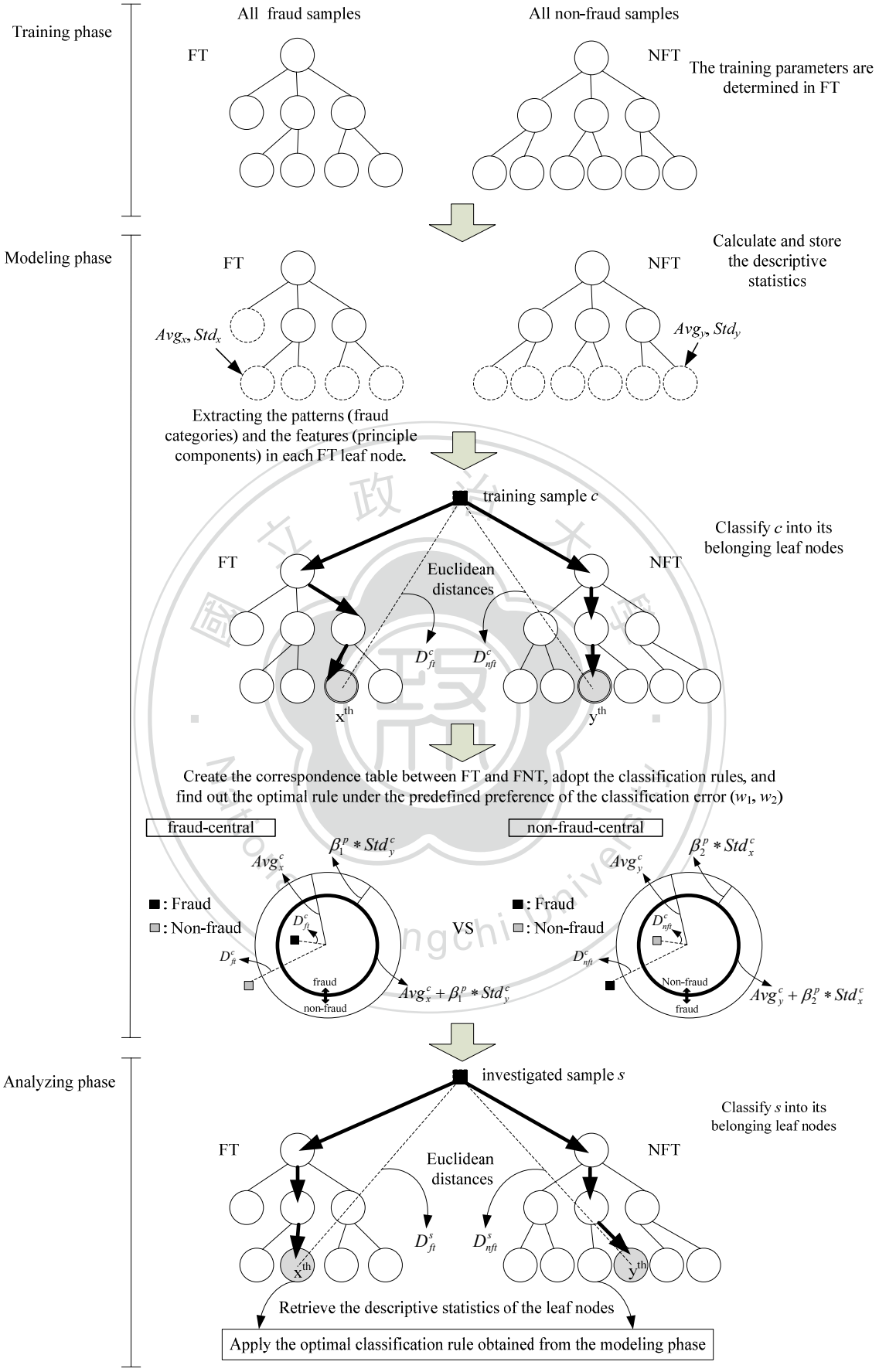


Figure 6. The classification concept of the proposed dual approach.

4. The FFR experiment and results

The FFR experiment takes a further step to identify whether there is a certain spatial relationship among fraud and non-fraud samples and, if it does, to derive the corresponding classification rule. Note that the proposed dual approach is data-driven which means that the corresponding system modeling is performed via using the sampled data. The details and the experimental results are briefed as follows.

4.1 Training phase – sampling module

The following sources are used to identify the fraud samples between the years from 1992 to 2006: indictments and sentences for major securities crimes issued by the Securities and Futures Bureau of the Financial Supervisory Commission, class action litigation cases initiated by Securities and Futures Investors Protection Center, and the law and regulations retrieving system of the Judicial Yuan in Taiwan. If a company's financial statement for a specific year is confirmed to be fraudulent by the indictments and sentences for major securities crimes issued by the Department of Justice, it is classified into our fraud observations. For those financial statements that are free from fraud allegations are classified into our non-fraud observations.

The matched-firm design is then used to form a sample set. That is, for each fraud firm, we match a non-fraud firm based on industry, total assets, and year. Thus, our sample composites of 116 publicly traded companies, including 58 fraud and 58 non-fraud ones over the period from 1992 to 2006. For each fraud company, we first identify the earliest year in which the financial statement fraud was committed. Then the sample periods cover two years before and two years after the year of the event. That is, five consecutive annual financial statements are used in our study. The final observations consist of 580 firm-year observations (i.e., annual financial statements) which comprise 113 fraud samples and 467 non-fraud samples. The sampling procedure is referred from Hsu (2008) and Huang et al. (2011)'s studies.

The firms are listed in Table 7. In addition, accounting rules, asset valuations and criteria governing preparation of financial statements for financial industry are incomparable with other industries so cases involving financial firms are excluded from the sample (Fanning and Cogger, 1998; Stice, 1991). Many literature (Beasley,

1996; Fanning and Cogger, 1998; Farber, 2005) based on sample of American firms used Accounting and Auditing Enforcement Releases (AAERs) issued by United States Securities and Exchange Commission (SEC) to determine whether or not firms committed financial reporting fraud. However, we had no such a consistent criterion in Taiwan, so the study established a criterion in the light of governmental publications and experts' opinion.

Table 7. The list of fraud and non-fraud firms in training samples.

No	fraud firm					non-fraud firm		Sampling period
	Industry	SIC Code	Name	Fraud year	Detect year	SIC Code	Name	
1	Electron	2398	博達	1999-2001	2004	3024	憶聲	1997-2001
2	Electron	8295	中強電	1998	1999	2349	錸德	1996-2000
3	Electron	2328	廣宇	1997-1998	1998	2411	飛瑞	1995-1999
4	Electron	2350	環隆電氣	1997-1998	1999	3037	欣興	1995-1999
5	Electron	2407	欣煜(前陞技)	2002-2004	2005	2316	楠梓電	2000-2004
6	Electron	2334	國豐	1997-1999	2001	2323	中環	1995-1999
7	Electron	2490	皇統科技	2000-2002	2004	2453	凌群	1998-2002
8	Electron	3039	宏傳	2004	2005	5353	台林	2000-2004
9	Electron	3001	協和國際	1999-2001	2004	8026	康和資	1997-2001
10	Electron	2494	突破	2002	2003	2419	仲琦	2000-2004
11	Electron	8188	麥瑟半導體	2001	2002	2425	承啟	1999-2003
12	Electron	6145	勁永國際	2003-2004	2005	8172	勝開	2001-2005
13	Electron	6250	宇加 (前太萊晶體)	2004	2005	3207	耀勝	2002-2006
14	Electron	5385	瑩寶科技	2000-2001	2002	5305	敦南	1998-2002
15	Iron& Steel	8708	大中鋼鐵	1997-1999	1999	2022	聚亨	1995-1999
16	Iron& Steel	2005	友力	1998-1999	1999	5009	榮剛	1996-2000
17	Iron& Steel	2019	桂宏	1998-2000	2000	2010	春源	1996-2000
18	Iron& Steel	2016	名佳利	1997-	1999	2032	新鋼	1995-1999

19	Iron& Steel	8714	紐新	1998 1997-1999	2001	2008	高興昌	1995-1999
20	Iron& Steel	5007	三星五金	1998-1999	2001	2009	第一銅	1996-2000
21	Iron& Steel	2017	官田鋼(前嘉益鋼)	1996-1998	2006	2013	中鋼構	1994-1998
22	Iron& Steel	8705	東隆五金	1997-1998	1998	9905	大華	1995-1999
23	Iron& Steel	2014	中鴻	2001-2003	2006	2006	東鋼	1999-2003
24	Building Material& Construction	1436	福益	1997-1998	1998	2535	達欣工	1995-1999
25	Building Material& Construction	2529	仁翔	1998	2001	2516	新建	1996-2000
26	Building Material& Construction	5503	榮美開發	2000	2003	2536	宏普	1996-2000
27	Building Material& Construction	2505	國揚實業	1997-1998	1998	2526	大陸	1995-1999
28	Building Material& Construction	8719	宏福	1997-1998	1999	2511	太子	1995-1999
29	Building Material& Construction	8716	尖美	1998-1999	2002	2524	京城	1995-1999
30	Building Material& Construction	2553	啟阜	1998-1999	2000	2534	宏盛	1996-2000
31	Building Material& Construction	5504	信南	1999-2000	2000	5514	三豐	1997-2001
32	Building Material& Construction	8710	易欣	1999	2000	5506	長鴻	1997-2001
33	Food	8723	順大裕	1998	1999	1201	味全	1996-2000
34	Food	8724	立大	1999-2000	2001	1219	福壽	1997-2001
35	Food	1221	久津	2001-2003	2003	1220	台榮	1999-2003
36	Food	1207	嘉食化	1998-2000	2006	1216	統一	1996-2000
37	Textile	1466	聚隆	1998	1999	1451	年興	1996-2000
38	Textile	8706	金緯	1998	1999	1423	利華	1995-1999

39	Textile	8717	瑞圓	1998	2001	1417	嘉裕	1995-1999
40	Electric Machinery	1505	楊鐵工廠	1997-1999	2001	1514	亞力	1997-2001
41	Plastic	8711	大穎	1999	2000	1304	台聚	1997-2001
42	Plastic	8713	延穎	1999	2000	1305	華夏	1997-2001
43	Electrical and Cable	1601	台光	1997-1998	1999	1614	三洋	1995-1999
44	Electrical and Cable	1602	太電	1994-1996	2003	1605	華新	1992-1996
45	Chemical	8701	正豐	1995-1996	1998	1716	永信	1993-1997
46	Chemical	4113	聯上生技	2004	2004	4123	晟德	2002-2006
47	Automobile	8712	國產車	1998	1998	2204	中華	1996-2000
48	Automobile	8702	羽田	1994-1995	1995	2201	裕隆	1992-1996
49	Automobile	2206	三陽工業	1999-2000	2001	2207	和泰	1997-2001
50	Shipping& Transportation	2614	東森國際	1999	2000	2615	萬海	1997-2001
51	Shipping& Transportation	2613	中櫃	1999	2000	5604	中連	1997-2001
52	Trading& Consumers' Goods	9801	力霸	1998-2000	2006	2903	遠百	1996-2000
53	Trading& Consumers' Goods	2913	農林	1996	1997	2915	潤泰全	1994-1998
54	Trading& Consumers' Goods	5901	中友百貨	1997-1999	2001	2905	三商行	1995-1999
55	Paper, Pulp	1918	萬有紙廠	1996-1998	1998	1902	台紙	1994-1998
56	Rubber	2101	南港輪胎	1997-1999	2001	2103	台橡	1995-1999
57	Other	8382	美式家具	1998	1999	9935	慶豐富	1996-2000
58	Other	9911	台灣櫻花	1998	2004	9915	億豐	1996-2000

Source: (Hsu, 2008).

Note that the financial statements in Taiwan are prepared according to International Financial Reporting Standards similar to the generally accepted accounting principles (GAAPs) adopted in the States, and the FFR fraud categories are identified with the COSO framework from Beasley et al. (1999), therefore, the proposed dual approach and findings of this study are generalizable.

4.2 Training phase – variable-selecting module

The independent variable is named FRAUD, which means if a company's financial statements for specific years are confirmed to be fraudulent by the indictments and sentences for major securities crimes issued by the Department of Justice, the firm-year data are classified into fraud observations, and the variable FRAUD will be set to 1, 0 otherwise. In terms of the independent variables, based upon literature regarding FFR, 25 explanatory variables are selected and are used to test the multi-collinearity effect before incorporated into the discriminant analysis. Table 8 (Hsu, 2008; Huang et al., 2012) summarizes the definition and measurement of these variables. These are measurement proxies for attributes of profitability, liquidity, operating ability, financial structure, cash flow ability, financial difficulty, and corporate governance[1] of a firm. These explanatory variables are collected from the Taiwan Economic Journal (TEJ) database.

The variables used by Persons (1995) are mostly measured on the basis of total assets. Persons (1995) concluded that financial leverage, asset composition and capital turnover are significant indicators in detection of fraudulent financial statements. Many research suggested that unethical managers often perpetrated frauds in accounts receivable and inventory because the account which involves subjective judgments increases audit risks, that is to say, auditors have difficulties to confirm validity of figures (American Institute of Certified Public Accountants 2002). Beasley et al. (1999) showed that 24% of fraud firms misstated inventory and 21% of fraud ones defraud in accounts receivable. In addition, Dechow et al. (2007) suggested that manipulation of accounts receivable improves sales growth and manipulation of inventory improves gross margin. Hence, this study used not only seven relevant input variables to measure a company's variation of inventory and accounts receivable but also several variables to observe profit and sales.

Significant declines in growth and profitability could put extreme pressures on management due to excessive expectations of third parties and risk of bankruptcy,

[1] Based on the governance characteristics of companies in East Asian economies suggested by Claessens et al. (2000), we employ *SMLSR* to proxy for ownership structure, *DBCRCFR* and *DBCBSCFR* for voting-right deviation (the difference between voting right and cash flow right) and seat-control deviation (the difference between the percentage of board seats controlled by the ultimate owners and cash flow right), and *SPR* for the risk dimension of the board members.

foreclosure, or hostile takeover (American Institute of Certified Public Accountants 2002). As a result, managers who especially experienced rapid growth attempt to practice deception to cover up crisis. It means that rapid sales or inconsistent profit is expected to be associated with the incidence of fraud (Bell and Carcello, 2000; Summers and Sweeney, 1998). This study employed net sales and net income to measure growth of a company.

Dechow et al. (2007) noted that change in free cash flows is a more fundamental measure than earnings because it abstracts from accruals. When accrual-based income statement is inconsistent with cash-based cash flow statement, it requires long-term observations. Hence, this study used two variables to measure adequacy and reinvestment percentage of cash flow over five years.

Fanning and Cogger (1998) suggested financial distress may be a motivation for management fraud. The managers who failed to stand heavy stress may cook the books to hide financial crisis from stakeholders. Loebbecke et al. (1989) found that 19% of fraud companies underwent solvency problems. On the other hand, the occurrence of financial crisis may result from weak corporate governance (Lee and Yeh, 2004). The fragile mechanism gives opportunities for managers to misrepresent easily and even frequently, hence this study investigated the relationship between financial statement fraud and corporate governance through four corporate governance indicators from the research of financial distress.

■ **Financial Ratios**

1. Profitability

(1) Gross profit margin (GPM): The GPM variable indicates a company ability to earn profits where the higher profit a company makes, the more unique competitive advantage a company owns. GPM can be defined as:

$$\frac{\text{Operating income} - \text{operating costs}}{\text{Operating income}}$$

(2) Operating profit ratio (OPR): The OPR variable usually reflects a company's profitability in its own industry. The difference between OPR and GPM is GPM only concerns direct costs from manufacturing products whereas OPR considers all costs in process of generating revenue. OPR can be defined as:

$$\frac{\text{Operating income} - \text{operating costs - operating expenses}}{\text{Operating income}}$$

(3) Return on assets (ROA): Persons (1995) indicated that lower profit may give management an incentive to overstate revenues or understate expenses. The ROA variable shows how much value a company's assets can carry in producing income before leverage. The higher ROA is, the better ability to utilize assets a company has. ROA can be defined as:

$$\frac{[(\text{Net income} + \text{Interest expenses} \times (1 - \text{tax rate})]}{\text{Average total assets}} - 1$$

(4) Growth rate of net sales (GRONS): The GRONS variable indicates a company's variation of sales revenues. GRONS can be defined as:

$$\left(\frac{\text{Net sales}}{\text{Net sales in prior fiscal year}}\right) - 1$$

(5) Growth rate of net income (GRONI): The GRONI variable indicates a company's variation of net income. GRONI can be defined as:

$$\left(\frac{\text{Net income}}{\text{Net income in prior fiscal year}}\right) - 1$$

2. Liquidity

(1) Current ratio (CR): The CR variable is to measure whether or not a company has enough current assets to pay short-term debts. The current assets is expected to be transformed into cash within one year, including cash equivalents, accounts receivable, prepaid expenses, inventory etc. A company with higher CR owns stronger ability to pay debts. CR can be defined as:

$$\frac{\text{Current assets}}{\text{Current liabilities}}$$

(2) Quick ratio (QR): The QR variable is to examine a company's ability to extinguish its short-term debts instantly. The inventory and prepaid expenses are excluded from quick assets. QR can be defined as:

$$\frac{\text{(Current assets - Inventories - prepaid expenses)}}{\text{Current liabilities}}$$

3. Operating ability

(1) Accounts receivable turnover (ART): The ART variable measures the frequency of accounts receivable collected during the period. ART relates to a company's efficiency of collection and adequacy of credit policy. ART can be defined as:

$$\frac{\text{Net credit sales}}{\text{Average accounts receivable}}$$

(2) Total asset turnover (TAT): The TAT variable indicates is used to determine how much sales revenue a company gains from investing in assets, in other words, a company's efficiency of utilizing its assets. Persons (1995) showed that managers of fraud companies may be incompetent to utilize assets to generating sales. TAT can be defined as:

$$\frac{\text{Net sales}}{\text{Total assets}}$$

(3) Growth rate of accounts receivable (GROAR): The GROAR variable indicates a company's variation of accounts receivable. GROAR can be defined as:

$$\left(\frac{\text{Accounts receivable}}{\text{Accounts receivable in prior fiscal year}}\right)-1$$

(4) Growth rate of inventory (GROI): The GROI variable indicates a company's variation of inventory. GROI can be defined as:

$$\left(\frac{\text{Inventory}}{\text{Inventory in prior fiscal year}}\right)-1$$

(5) Growth rate of Accounts receivable to gross sales (GRARTGS): The GRARTGS variable indicates a company's variation of ratio of accounts receivable to gross sales. GRARTGS can be defined as:

$$\frac{\text{Accounts receivable}_t}{\text{Gross sales}_t} - \frac{\text{Accounts receivable}_{t-1}}{\text{Gross sales}_{t-1}}$$

(6) Growth rate of Inventory to gross sales (GRITGS): The GRITGS variable indicates a company's variation of ratio of Inventory to gross sales. GRITGS can be defined as:

$$\frac{\text{Inventory}_t}{\text{Gross sales}_t} - \frac{\text{Inventory}_{t-1}}{\text{Gross sales}_{t-1}}$$

(7) Accounts receivable to total assets (ARTTA): Persons (1995) suggested the current assets of fraud firms consist mostly of receivables and inventories, so the study used two variables: ARTTA and ITTA to determine a firm's asset composition. The ARTTA variable is used to examine the percentage of accounts receivable in total assets. ARTTA can be defined as:

$$\frac{\text{Accounts receivable}}{\text{Total assets}}$$

(8) Inventory to total assets (ITTA): The ITTA variable is used to examine the percentage of inventory in total assets. ITTA can be defined as:

$$\frac{\text{Inventory}}{\text{Total assets}}$$

4. Financial structure

(1) Debt ratio (DR): The DR variable is used to measure a company's capital structure and financial leverage. The debt financing not only raises return on investment, but also has benefits of tax shield substitute. But higher leverage increase risk of bankruptcy, Persons (1995) found that fraud firms have higher financial leverage than non-fraud firms. DR can be defined as:

$$\frac{\text{(Total liabilities)}}{\text{Total assets}}$$

(2) Long-term funds to fixed assets (LFTFA): The LFTFA variable is used to measure the degree of fixed assets provided by long-term funds. Higher LFTFA means the capital structure of a company may be sound because investment in fixed assets usually require long collection period. LFTFA can be defined as:

$$\frac{(\text{Equity} + \text{longterm liabilities})}{\text{Fixed assets}}$$

5. Cash flow ability

(1) Cash flow ratio (CFR): The CFR variable is used to assess a company's ability of paying current debts by cash. CFR differs from CR or QR in the measurement duration. CFR is determined by cash flows from operating activities of one fiscal year, not by one point. CFR can be defined as:

$$\frac{\textit{Cash flows from operating activities}}{\textit{Current liabilities}}$$

(2) Cash flow adequacy ratio (CFAR): The CFAR variable is used to evaluate whether or not cash flows from operating activities is enough to disburse in capital expenditures, inventory and cash dividends. CFAR can be defined as:

$$\frac{\text{Five- year sum of cash flows from operating activities}}{\text{Five- year sum of capital expenditures, inventory additions and cash dividends.}}$$

(3) Cash flow reinvestment ratio (CFRR): The CFRR variable is used to determine the percentage of utilizing cash flows from operating activities to reinvest in assets and firm development. CFRR can be defined as:

$$\frac{\text{(Cash flows from operating a activities- cash dividends)}}{\text{Gross fixed assets + long- term investments+ other assets + working capital}}$$

■ **Corporate Governance Indicator**

1. Stock Pledge ratio (SPR): This study employed SPR variable to determine whether financial distress happens or not. SPR variable means the percentage of shareholdings which directors and supervisors put in pledge for loans and credits. Directors and supervisors often pledge their stocks to obtain funds to keep stock price as well as rescue firms from financial distress (Lee and Yeh

2004). Nevertheless, high personal leverage and excessive investment in stock market could expose company to financial risk. SPR can be defined as:

$$\frac{\text{Sahresholdings in pledge}}{\text{Total shareholdings}}$$

2. Shareholdings of major shareholders ratio (SOMSR): The research employed SOMSR variable to assess independence of the board. Excluding directors, supervisors or managers, major shareholders are defined as shareholders whose percentage of shareholdings is greater than 10% according to Taiwan Stock Exchange (TSE). The SOMSR variable means the sum of percentage of major shareholders' shareholdings. The major shareholders who own higher percentage of shareholdings have much motivation to supervise managers. It could reduce not only agency problem, but also the probability of financial distress.

3. Deviation between control rights and cash flow rights (DBCRCFR): The study adopted the indicator to measure integrity of corporate governance. Lee and Yeh (2004) noted that the larger difference between voting and cash flow rights is, the stronger incentive to expropriate minority interests ultimate owners have. It may result in malfeasance or even financial distress. DBCRCFR can be written as:

DBCRCFR＝Control rights－Cash flow rights

4. Deviation between ratio of controlled board seats and cash flow rights (DBCBSCFR): The study used the variable to examine integrity of corporate governance. The larger deviation between percentage of controlled board seats and cash flow rights is, the more effortless to perpetrate misrepresentation or misappropriation controlling shareholders are. In other words, it not only puts a firm at financial risk, but also appear failures of scrutiny. DBCBSCFR can be written as:

DBRCBSCFR＝Percentage of Controlled board seats－Cash flow rights

(1) Control rights: The control rights, also called voting rights, indicates the shareholdings of ultimate owners who can greatly influence corporate decision according to the definition of La Porta, Lopez-de-Silanes, and

Shleifer (1999). The ultimate owners who usually involve major shareholders, chairman of board or management/family groups control the firm directly (i.e., through shares registered in his name) and indirectly (i.e., through shares held by entities that he controls). Control rights can be written as:

Control rights＝Direct control rights＋Indirect control rights

(2) Cash flow rights: The cash flow rights, also called earnings distribution rights, mean that ultimate owners gain earnings through direct shareholdings and indirect ownership, for instance, using cross-shareholdings. The indirect cash flow rights is summed up the product of successive ownership along every control chain (Lee and Yeh, 2004). Cash flow rights can be written as:

Cash flow rights＝Direct cash flow rights＋Indirect cash flow rights

According to the definition of Claessens, Djankov, and Lang (2000), the study suppose, for example, that a family owns 40% of stock of listed firm A and 20% of stock of listed firm B. Additionally, they acquire 30% of stock of firm A through firm B. When major shareholders don't exist in firm A and firm B, the family is the ultimate owners of both firms. The family controls directly 40% of firm A namely direct control right. The indirect control right which the family control firm A is chose the minimum between shareholdings of firm B and shareholdings through firm B. For this reason, we would say that the family owns totally 60% of control rights of firm A.

In terms of cash flow right, the family controls directly 40% of firm A namely direct cash flow right, but indirect cash flow right differs from the method of indirect control right. The indirect cash flow right is the product of shareholdings along each control chain, that is 20%×30%＝6%. Consequently, the family owns 46% of total cash flow rights of firm A.

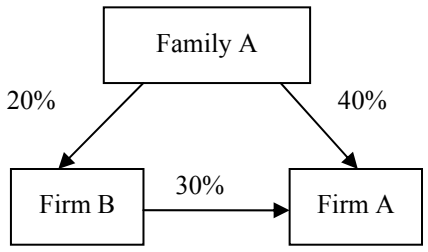


Figure 7. An example of control right and cash flow right.

(3) Percentage of controlled board seats: The variable indicates the number of board seats, including directors' and supervisors', are held by ultimate owners along the control chains. It appears the degree of control over the family or insiders. Yeh, Lee, and Woidtke (2001) also argued that the correlation between percentage of controlled board seats and firm financial performance is negative. Percentage of controlled board seats can be written as:

$$\frac{\text{Board seats held by conrolling sahresholders}}{\text{Total board seats}}$$

■ **Z-score**

The study used Altman Z scores (Altman, 1968) to measure a company's financial condition to determine the relationship between financial distress and fraud. The smaller Z score indicated that a firm may fail or go into bankruptcy with higher probability. Altman's Z score can be computed as:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

Where:

X_1=working capital/ total assets

X_2=retained earnings/ total assets

X_3=earnings before interest and taxes/ total assets

X_4=market value of equity/ book value of total debt

X_5=net sales/ total assets

Table 8. Variable definition and measurement.

Variable Definition	Literature	Measurement
Dependent variable:		
FRAUD	Persons (1995)	If a company's financial statements for specific years are confirmed to be fraudulent by the indictments and sentences for major securities crimes issued by the Department of Justice, the firm-year data are classified into fraud observations, and the variable *FRAUD* will be set to 1, 0 otherwise.
Independent variable		
Profitability		
GPM	Dechow et al. (2007)	$\frac{Sales - Operating\ costs}{Sales}$
OPR	Green and Choi (1997)	$\frac{Sales - Operating\ costs - Operating\ expenses}{Sales}$
ROA	Persons (1995), Hoogs et al. (2007)	$\frac{Net\ income + Interest\ expenses \times (1 - Tax\ rate)}{Average\ total\ assets}$
GRONS	Stice (1991), Summers and Sweeney (1998), Dechow et al. (2007)	$(\frac{Sales}{Sales\ in\ prior\ fiscal\ year}) - 1$
GRONI		$(\frac{Net\ income}{Net\ income\ in\ prior\ fiscal\ year}) - 1$
Liquidity		
CR	Kirkos et al. (2007)	$\frac{Current\ assets}{Current\ liabilities}$
QR		$\frac{Current\ assets - Inventories - Prepaid\ expenses}{Current\ liabilities}$
Operating ability		
ART	Green and Choi (1997)	$\frac{Sales}{Average\ accounts\ receivable}$
TAT	Persons (1995), Kirkos et al. (2007)	$\frac{Sales}{Total\ assets}$
GROAR	Dechow et al. (2007)	$(\frac{Accounts\ receivable}{Accounts\ receivable\ in\ prior\ fiscal\ year}) - 1$
GROI		$(\frac{Inventory}{Inventory\ in\ prior\ fiscal\ year}) - 1$

GRARTGS	Summers and Sweeney (1998)	$\frac{Accounts\ receivable_t}{Sales_t} - \frac{Accounts\ receivable_{t-1}}{Sales_{t-1}}$
GRITGS		$\frac{Inventory_t}{Sales_t} - \frac{Inventory_{t-1}}{Sales_{t-1}}$
ARTTA	Stice (1991), Persons (1995), Green and Choi (1997)	$\frac{Accounts\ receivable}{Total\ assets}$
ITTA		$\frac{Inventory}{Total\ assets}$
Financial structure		
DR	Persons (1995), Kirkos et al. (2007)	$\frac{Total\ liabilities}{Total\ assets}$
LFTFA		$\frac{Equity + Longterm\ liabilities}{Fixed\ assets}$
Cash flow ability		
CFR		$\frac{Cash\ flows\ from\ operating\ activities}{Current\ liabilities}$
CFAR	Dechow et al. (2007)	$\frac{Five\ year\ sum\ of\ cash\ flows\ from\ operating\ activities}{(Five\ year\ sum\ of\ capital\ expenditures,\ inventory\ additions\ and\ cash\ dividends)}$
CFRR		$\frac{Cash\ flows\ from\ operating\ activities - Cash\ dividends}{(Gross\ fixed\ assets + Long\ term\ investments + Other\ assets + Working\ capital)}$
Financial difficulty		
Z-score	Altman (1968), Stice (1991), Summers and Sweeney (1998), Fanning and Cogger (1998)	$1.2\times(\frac{Working\ capital}{Total\ assets}) + 1.4\times(\frac{Retained\ earnings}{Total\ assets}) + 3.3\times(\frac{Earnings\ before\ interest\ and\ taxes}{Total\ assets}) + 0.6\times(\frac{Market\ value\ of\ equity}{Book\ value\ of\ total\ debt}) + 1.0\times TAT$
Corporate Governance		
SPR[#]	Lee and Yeh (2004)	$\frac{large\ shareholders'\ shareholdings\ in\ pledge}{large\ shareholders'\ shareholdings}$
SOMSR	Beasley et al. (1999)	Σ (*Percentage of shareholdings >10%*)

DBCRCFR	La Porta et al. (1999), Lee and Yeh (2004)	*Voting rights (CR) - Cash flow rights (CFR)*
DBCBSCFR	Yeh et al. (2001)	*Percentage of board seats controlled (CBS)- Cash flow rights (CFR)*

[#]: According to the rule issued by the Securities and Futures Commission (SFC) of Taiwan, for public companies, the board members, managers and major shareholders (who own 10 percent or more of a company's outstanding shares) of the company are obliged to report to the SFC the percentage of their shareholdings been pledged for loans and credits. (Lee and Yeh, 2004)

The results of the multi-collinearity test indicate that one variable - *GRITGS* - should be excluded. As a result, 24 independent variables are kept that will be used as the input variables for the GHSOM. These are measurement proxies for attributes of profitability, liquidity, operating ability, financial structure, cash flow ability, financial difficulty, and corporate governance of a firm. These explanatory variables are collected from the Taiwan Economic Journal (TEJ) database. The variable selection procedure is referred from Hsu (2008) and Huang et al. (2011)'s studies.

Table 9 reports the empirical results of the discriminant analysis (Hsu 2008; Huang et al., 2011). The analysis that the Wilks' Λ value equals to 0.766 and x^2 equals to 151.095 (both significant at p-value < 0.01) suggests that the discriminant model employed has adequate explanatory power. The results of discriminant analysis indicate that eight variables, return on assets (ROA), current ratio (CR), quick ratio (QR), debt ratio (DR), cash flow ratio (CFR), cash flow adequacy ratio (CFAR), Z-Score and sock pledge ratio (SPR), are significant at p-value < 0.01 level. These eight variables are collected for our sample firms and used as the training data for the GHSOM.[2] These eight variables proxy a company's attributes from the aspects of profitability (ROA), liquidity (CR and QR), financial structure (DR), cash flow ability (CFR and CFAR), financial difficulty (Z-Score), and corporate governance (SPR).

[2] We have also performed the logistic regression in the data preprocessing stage and the results indicate that there are only two variables, *ROA* and *CFR*, are significant at p-value < 0.01 level. Although the number of input variables resulted from the data preprocessing does affect the implementation efficiency of GHSOM, the performance of data preprocessing in any application of GHSOM has a nature of the exploratory data analysis and its purpose is to form a set of candidate variables for GHSOM. In order to form a set of candidates for GHSOM, rather than a set of significant predictors to a linear prediction model (such as logistic regression or discriminant analysis), we take a union of these two sets, which is the same set from the discriminant analysis.

Table 9. Empirical results of discriminant analysis.

Variable	Coefficient	F-value	Significance
GPM	0.14	3.51	0.061
OPR	-0.03	0.16	0.688
ROA	0.77***	105.82	0.000
GRONS	0.06	0.63	0.427
GRONI	-0.02	0.05	0.822
CR	0.34***	20.59	0.000
QR	0.28***	13.42	0.000
ART	0.09	1.58	0.210
TAT	0.19	6.38	0.012
GROAR	0.03	0.12	0.731
GROI	0.07	0.90	0.344
GRARTGS	0.00	0.00	0.997
ARTTA	0.11	2.25	0.134
ITTA	0.12	2.37	0.125
DR	-0.42***	30.46	0.000
LFTFA	0.02	0.09	0.764
CFR	0.33***	19.21	0.000
CFAR	0.24***	9.89	0.002
CFRR	0.19	6.41	0.012
SPR	-0.47***	38.85	0.000
SOMSR	-0.19	6.18	0.013
DBCRCFR	0.02	0.04	0.835
DBCBSCFR	-0.05	0.41	0.524
Z-score	0.64***	72.74	0.000
Wilks' Λ value	0.77	p-value	0.000
χ^2	151.10	p-value	0.000

*** p-value significant at <0.01 level.

4.3 Training phase – clustering module

As stated in (Dittenbach et al., 2000), the development of the GHSOM is primarily dominated by the parameters of breadth (τ_1) and depth (τ_2). In order to reach the goal of

obtaining the multi-layer hierarchy feature and preventing the overly clustering of fraud samples, we predefined the following selection criteria to derive an acceptable FT:

1) There are more than one layers of SOM in the GHSOM.
2) Samples of each node should not be overly clustered, and each leaf node should at least contain one sample.
3) The number of subgroups and the corresponding value of mean quantitative error (MQE) among nodes are considered for evaluating the clustering result. The MQE value indicates the samples homogeneity among clusters; the clustering quality is better if the MQE value is smaller. The clustering method that leads to a smaller number of subgroups and a lower MQE value is better.

Based on the criteria aforementioned, the trials of the GHSOM parameter setting are taken and shown in Table 10. The parameter τ_1 is adjusted from 0.5 to 0.8 per 0.1 scales, and the parameter τ_2 is adjusted from 0.05 to 0.07 per 0.01 scales. When $\tau_1 = 0.6$ and $\tau_2 = 0.07$, each leaf node has at least one fraud sample. In the condition of same MQEs, the parameter setting $\tau_1 = 0.8$ and $\tau_2 = 0.07$ leads to less number of leaf nodes. Therefore, the parameter setting $\tau_1 = 0.8$ and $\tau_2 = 0.07$ are used to generate FT and NFT, respectively.

Table 10. The GHSOM parameter setting trials.

Breadth	Depth	layer	leaves	MQE	each group exists at last one sample
0.5	0.05	3	28	0.014091	no
0.6	0.05	3	20	0.024047	no
0.7	0.05	3	18	0.024047	no
0.8	0.05	3	16	0.024047	no
0.5	0.06	3	28	0.014091	no
0.6	0.06	3	20	0.024047	no
0.7	0.06	3	18	0.024047	no
0.8	0.06	3	16	0.024047	no
0.5	0.07	3	20	0.014091	no
0.6	0.07	3	17	0.024047	yes
0.7	0.07	3	15	0.024047	yes
0.8*	0.07	3	13	0.024047	yes

* The chosen GHSOM tree

We use the GHSOM toolbox in the platform of Matlab R2007a to generate FT and NFT. The obtained GHSOMs are shown in Figure 8. The leaf nodes are marked in taint. For each node, a name in numerical label is given according to its layer number and its node order in the same SOM as well as its parent's name. For instance, the node #13-24 is node number 4 in layer 2 developed from the node number 3 in layer 1 of FT. Based on the clustering result, we believe that it is plausible to extract the distinctive (common) patterns or features of these leaf nodes.

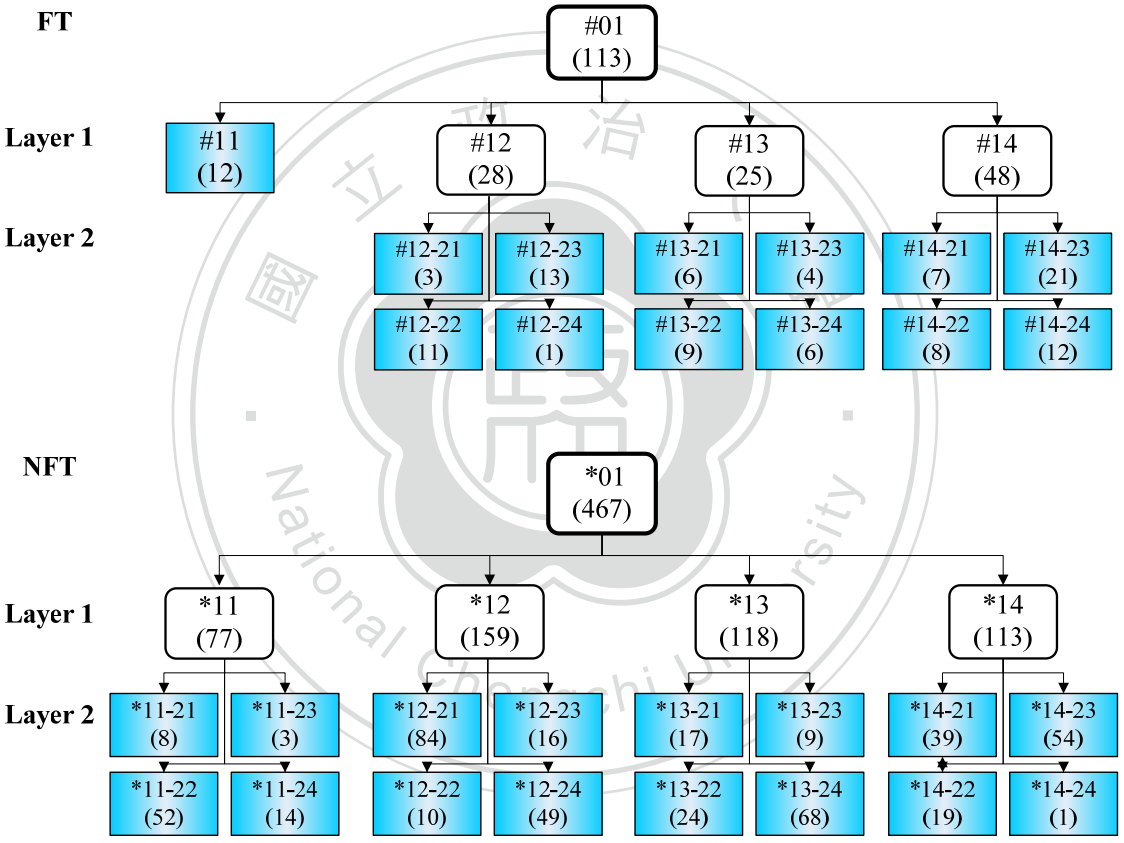


Figure 8. The obtained FT and NFT.

As shown in Figure 8, the FT and NFT have different GHSOM structures since that the leaf node *11 of NFT generates four child nodes while the leaf node #11 of FT does not grow further. Note that the names and orders of leaf nodes of FT and NFT do not release any spatial implication among them.

4.4 Modeling phase – statistic-gathering, rule-forming module

The leaf node matching from NFT to FT via all (fraud and non-fraud) training samples is shown in the first three columns of Table 11 and Figure 9. For example, the leaf node #12-21 of FT hosts 80% of the training samples classified into the leaf node *11-21 of NFT. That is, there are 80% of the training samples in #12-21 classified into *11-21. Hence, the leaf node #12-21 of FT is matched to the leaf node *11-21 of NFT and claim the leaf node #12-21 of FT is the counterpart of the leaf node *11-21 of NFT. That is, the fraud samples classified into the leaf node #12-21 of FT cluster around the non-fraud samples in the leaf node *11-21 of NFT. Based on the proportional majority, the counterparts of a leaf node of NFT could be more than one. For example, the leaf nodes #12-23 and #12-22 of FT host 93.33% of the training samples classified into the leaf node *11-24 of NFT. Thus, the leaf nodes #12-23 and #12-22 of FT are the counterparts of the leaf node *11-24 of NFT. The corresponding Avg and Std values of NFT_y and FT_x are shown in the fourth and the fifth columns of Table 11.

Table 11. The leaf node matching from NFT to FT.

*NFT	#FT	Samples proportion	Avg_y of NFT	Std_x of FT	Classification error (%)*
11-21	12-21	80.00%	0.65	0.83	(0.00, 20.00)
11-22	12-23	86.15%	0.23	0.32	(7.69, 15.38)
11-24	12-23	73.33%	0.29	0.32	(0.00, 6.67)
	12-22	20.00%	0.29	0.19	
12-21	12-22	100.00%	0.18	0.35	(0.00, 9.68)
12-22	12-22	100.00%	0.79	0.35	(9.09, 0.00)
12-23	12-22	100.00%	0.30	0.35	(0.00, 0.00)
12-24	12-22	100.00%	0.27	0.35	(0.49, 3.92)
13-21	11	79.17%	0.39	0.73	(0.00, 16.67)
13-22	14-21	45.83%	0.31	0.25	(14.29, 26.53)
	11	33.33%	0.31	0.73	
	13-24	14.58%	0.31	0.35	
13-23	13-21	76.92%	0.54	1.08	(6.67, 33.33)
	13-24	23.08%	0.54	0.35	
13-24	14-23	80.41%	0.30	0.26	(15.96, 23.4)
14-21	14-22	76.74%	0.26	0.27	(22.5, 7.5)
	12-22	16.28%	0.26	0.35	

14-22	14-24	65.00%	0.37	0.47	(20.00, 5.00)
	14-22	35.00%	0.37	0.27	
14-23	14-22	59.09%	0.28	0.27	(15.15, 12.12)
	14-24	37.88%	0.28	0.47	

* The numbers within the parenthesis indicate the type I error and the type II error.

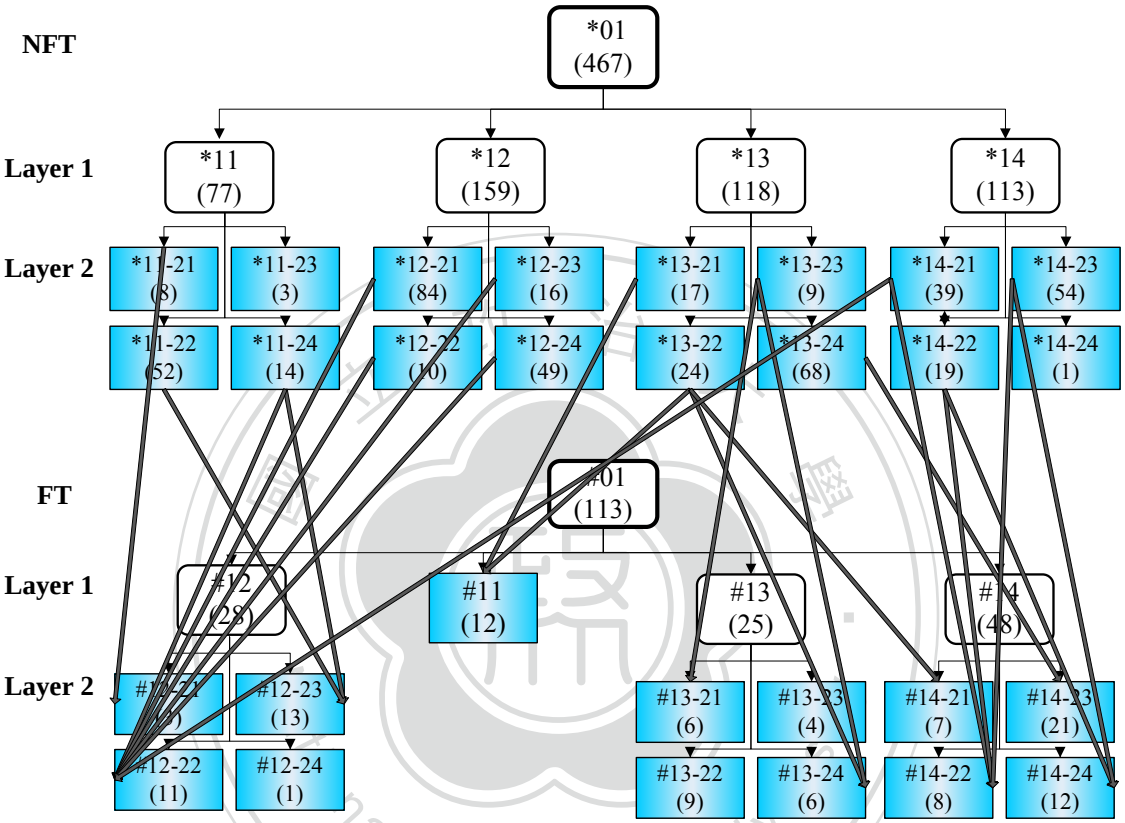


Figure 9. The leaf node matching from NFT to FT.

Following the optimization approach stated in Section 3, the parameter β_2 of the non-fraud-central rule is tuned by solving the optimization problem (8) regarding each of the three different settings of the constants w_1 and w_2. The corresponding classification errors (type I and type II) are stated in the last column of Table 12. For example, the setting of w_1 = 0.01 and w_2 = 1 for the match leaf nodes *11-22 and #12-23 has the smallest corresponding sum of (type I and type II) classification errors and the corresponding optimal β_2 values are within the range of [0.1531, 0.1562]. Later we set β_2 = 0.153 for samples classified to the leaf node *11-22. As shown in Table 11, the corresponding type I error is 13.62% and type II error is 13.28% regarding all 580 training samples.

Table 12. The result of w_1 and w_2 of the non-fraud-central rule.

*NFT→#FT	w_1	w_2	β_2	Classification error (%)*
*11-21→#12-21	1	1	0.3976 ~0.4172	(0.00%, 7.14%)
*11-22→#12-23	0.01	1	0.1531 ~0.1562	(8.77%, 17.54%)
*11-24→#12-23 #12-22	1	0.01	1.9114 ~3.937	(0.00%, 7.14%)
*12-21→#12-22	0.01	1	0.1531 ~0.1562	(15.05%, 7.53%)
*12-22→#12-22	0.01	1	0.1531 ~0.1562	(10.00%, 0.00%)
*12-24→#12-22	1	0.01	1.9114 ~3.937	(0.00%, 3.92%)
*13-21→#11	0.01	1	0.1531 ~0.1562	(0.00%, 15.79%)
*13-22→#14-21 #11 #13-24	0.01	1	0.1531 ~0.1562	(15.56%, 22.22%)
*13-23→#13-21 #13-24	0.01	1	0.1531 ~0.1562	(7.69%, 30.77%)
*13-24→#14-23	0.01	1	0.1531 ~0.1562	(16.13%, 22.58%)
*14-21→#14-22 #12-22	1	0.01	1.9114 ~3.937	(1.82%, 27.27%)
*14-22→#14-24 #14-22	0.01	1	0.1531 ~0.1562	(11.43%, 14.29%)
*14-23→#14-22 #14-24	0.01	1	0.1531 ~0.1562	(13.41%, 10.98%)

* The numbers within the parenthesis indicate the type I error and the type II error, respectively.

In contrast, the leaf node matching from FT to NFT via all (fraud and non-fraud) training samples is shown in the first three columns of Table 13 and Figure 10. The corresponding Avg and Std values of FT_x and NFT_y are shown in the fourth and the fifth columns of Table 13.

Table 13. The leaf node matching from FT to NFT.

#FT	*NFT	Sample proportion	Avg_x of FT	Std_y of NFT	Classification error (%)*
11	13-21	43.18%	0.59	0.18	(4.55, 81.25)
	13-22	36.36%	0.59	0.14	
12-21	11-21	88.89%	0.32	0.13	(0.00, 100.00)
12-22	12-21	49.21%	0.19	0.08	(0.00, 100.00)
	12-24	26.98%	0.19	0.15	
12-23	11-22	83.58%	0.28	0.12	(7.46, 84.62)
13-21	13-23	66.67%	0.43	0.27	(28.57, 16.67)
	13-21	20.00%	0.43	0.18	
13-24	13-22	70.00%	0.32	0.14	(50.00, 40.00)
	13-23	30.00%	0.33	0.27	
14-21	13-22	61.11%	0.25	0.14	(30.56, 58.33)
	13-24	36.11%	0.25	0.17	
14-22	14-21	40.74%	0.24	0.12	(0.00, 100.00)
	14-23	48.15%	0.24	0.15	
14-23	13-24	95.12%	0.23	0.17	(29.63, 25.00)
14-24	14-23	58.14%	0.32	0.15	(0.00, 100.00)
	14-22	30.23%	0.32	0.17	

* The numbers within the parenthesis indicate the type I error and the type II error.

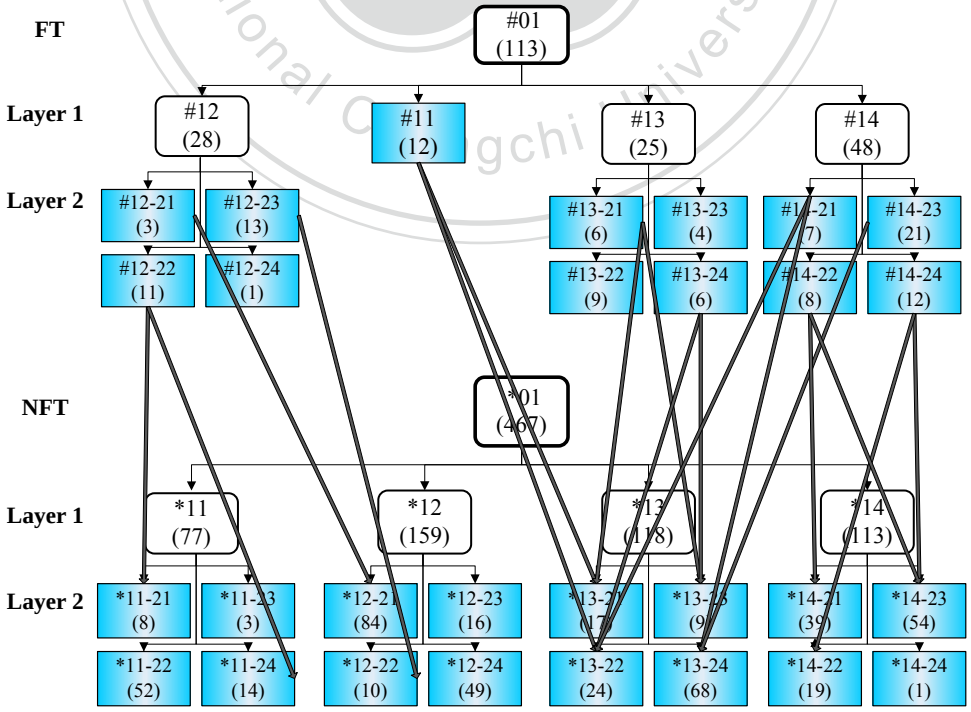


Figure 10. The leaf node matching from FT to NFT.

Following the optimization approach stated in Section 3, the parameter β_1 of the fraud-central rule is tuned by solving the optimization problem (4) regarding each of the three different settings of the constants w_1 and w_2. The corresponding classification errors (type I and type II) are stated in the last column of Table 14. For example, the setting of $w_1 = 0.01$ and $w_2 = 1$ for the match leaf nodes #12-21 and *11-21 has the smallest corresponding sum of (type I and type II) classification errors and the corresponding optimal β_1 values are within the range of [1.44, 1.447]. Later we set $\beta_1 = 1.44$ for samples classified to the leaf node #12-21. As shown in Table 13, the corresponding type I error is 8.78% and type II error is 76.11% regarding all 580 training samples.

Table 14. The result of w_1 and w_2 of the fraud-central rule.

#FT→*NFT	w_1	w_2	β_1	Classification error (%)*
#11→*13-21 *13-22	1	1	-0.293~-0.25	(2.86%, 34.29%)
#12-21→*11-21	0.01	1	1.44~1.447	(0.00%, 25.00%)
#12-22→*12-21 *12-24	1	0.01	-0.47~-0.42	(0.00%, 7.64%)
#12-23→*11-22	0.01	1	1.44~1.447	(8.93%, 17.86%)
#13-21→*13-23 *13-21	0.01	1	1.44~1.447	(30.77%, 7.69%)
#13-24→*13-22 *13-23	0.01	1	1.44~1.447	(50.00%, 20.00%)
#14-21→*13-22 *13-24	0.01	1	1.44~1.447	(31.43%, 20.00%)
#14-22→*14-21 *14-23	1	0.01	-0.47~-0.42	(0.00%, 13.51%)
#14-23→*13-24	0.01	1	1.44~1.447	(30.00%, 7.50%)
#14-24→*14-23 *14-22	1	1	-0.293~-0.25	(0.00%, 20.00%)

In sum, the non-fraud-central rule is better through comparing the corresponding sum of (type I and type II) classification errors in Table 12 with the ones in Table 14.

4.5 Modeling phase – feature-extracting module

Without loss of generalization, the clustering result of the GHSOM is used to illustrate the operation of a feature-extraction stage and we demonstrate the features of each leaf node of FT. The leaf node #12-24 is excluded due to having only one sample. For each leaf node of the GHSOM, values of the eight significant variables regarding all clustered samples are the inputs of PCA. According to Kaiser (1960), only those factors whose variances are greater than 1[3] are retained as the principle components. Table 15 presents the estimated eigenvalues of eight factors regarding all leaf nodes. According to the factor selection criterion, for instance, #11 has retained the first three factors as its principle components, in which factor 1 explains 44.819% of the total variance of the input variables, factor 2 27.842% and factor 3 15.964%.

Table 15. The estimated eigenvalues of eight factors regarding all FT leaf nodes.

Leaf node	Factor	Eigenvalue	% of Variance
#11	1	3.586	44.819
	2	2.227	27.842
	3	1.277	15.964
	4	0.423	5.292
	5	0.253	3.162
	6	0.138	1.723
	7	0.064	0.800
	8	0.032	0.398
#12-21	1	6.468	92.400
	2	0.532	7.600
	3	0.000	0.000
	4	0.000	0.000
	5	0.000	0.000
	6	0.000	0.000
	8	0.000	0.000
#12-22	1	2.691	38.440
	2	1.697	24.237
	3	1.253	17.905

[3] That is its corresponding eigenvalue is large than 1.

	4	0.802	11.458
	5	0.495	7.069
	6	0.039	0.563
	8	0.023	0.329
#12-23	1	3.618	51.689
	2	1.885	26.931
	3	0.953	13.608
	4	0.321	4.593
	5	0.151	2.161
	6	0.042	0.603
	8	0.029	0.415
#13-21	1	3.699	46.242
	2	1.923	24.038
	3	1.440	17.995
	4	0.822	10.271
	5	0.116	1.455
	6	0.000	0.000
	7	0.000	0.000
	8	0.000	0.000
#13-22	1	2.845	35.560
	2	2.633	32.919
	3	1.534	19.178
	4	0.629	7.860
	5	0.262	3.274
	6	0.055	0.684
	7	0.030	0.379
	8	0.012	0.147
#13-23	1	3.926	49.076
	2	3.196	39.949
	3	0.878	10.975
	4	0.000	0.000
	5	0.000	0.000
	6	0.000	0.000
	7	0.000	0.000
	8	0.000	0.000
#13-24	1	3.541	44.257
	2	2.883	36.044

	3	1.206	15.070
	4	0.257	3.219
	5	0.113	1.410
	6	0.000	0.000
	7	0.000	0.000
	8	0.000	0.000
#14-21	1	3.092	38.655
	2	2.589	32.361
	3	1.244	15.545
	4	0.850	10.626
	5	0.176	2.204
	6	0.049	0.609
	7	0.000	0.000
	8	0.000	0.000
#14-22	1	3.271	40.887
	2	1.846	23.079
	3	1.469	18.368
	4	0.765	9.562
	5	0.443	5.532
	6	0.172	2.153
	7	0.034	0.419
	8	0.000	0.000
#14-23	1	2.879	35.982
	2	2.100	26.247
	3	1.137	14.214
	4	0.766	9.580
	5	0.439	5.483
	6	0.333	4.166
	7	0.221	2.761
	8	0.125	1.567
#14-24	1	4.048	50.601
	2	2.027	25.336
	3	0.931	11.637
	4	0.660	8.251
	5	0.151	1.886
	6	0.114	1.425
	7	0.063	0.791

	8	0.006	0.074

Note: The values of the sixth factor (SPR) are the same in leaf nodes #12-21, #12-22 and #12-23. Therefore, they do not have eigenvalues.

Table 15 presents the estimated eigenvalues of eight factors regarding all leaf nodes of FT. The leaf node #13-21, #13-22, and #13-24 has three factors in which the eigenvalue is greater than 1. The leaf node #13-23 has two factors in which the eigenvalue is greater than 1. The leaf node #14-21, #14-22, and #14-23 has three factors whose eigenvalues are bigger than 1. The leaf node #14-24 has two principle factors whose eigenvalues are bigger than 1. Those factors with eigenvalues bigger than 1 are determined as the principle components of its belonging leaf node.

To enhance the interpretability of the obtained principle components, the varimax factor rotation method is used here. This method minimizes the number of variables that have high loadings of a principle component. To differentiate features in each principle component, variables with the absolute value of corresponding factor loadings less than 0.6 are omitted. Table 16 to Table 18 shows the results of a varimax factor rotation method regarding the leaf nodes of FT.

Table 16 shows the results of a varimax factor rotation method regarding all FT leaf nodes.

Table 16. The factor loadings of all FT leaf nodes.

Leaf node	Principle component	ROA	CR	QR	DR	CFR	CFAR	SPR	Z-score
	1		0.911						
#11	2			0.786		0.732	0.9		
	3	0.859						0.908	
#12-21	1	0.983	0.95	0.975	0.999	-0.93	-0.94		-0.949
	1				-0.863				0.797
#12-22	2					0.898			
	3	0.941					-0.778		
#12-23	1	0.706			-0.967				0.905
	2		-0.891			0.87	0.953		
#13-21	1		-0.902	0.821			0.95		
	2	0.886			-0.969				
	3							0.909	-0.967
#13-22	1			0.873	0.934				-0.892
	2					0.947	0.927		
	3		0.708					0.871	
#13-23	1	0.99	0.967	0.935				-0.816	
	2				-0.728	0.815	0.944		-0.999
#13-24	1		-0.929	-0.86		0.97	0.902		
	2	0.946			-0.931				0.891
	3							0.984	
#14-21	1					0.869	0.915	0.729	
	2		0.961	0.768	0.763				
	3								0.955
#14-22	1			0.63		0.969	0.995		
	2				-0.915				0.825
	3	0.774	0.648					0.828	
#14-23	1		0.668		-0.87		0.73		0.92
	2	0.74		0.83					
	3					0.886			
#14-24	1	0.851				0.979	0.968		0.957
	2		0.892	0.847	-0.645				

As shown in Table 16, the principle components extracted from different leaf nodes have a heterogeneous composite of variables. For instance, regarding the leaf node #11, its first principle component consists of one debt related ratio (CR); its second principle component consists of three liquidity related ratios (QR, CFR and CFAR); and its third principle component consists of one earning related and one corporate governance ratios (ROA and SPR). Hence, the first principle component represents debt paying ability of a firm; the second principle component represents the liquidity of a firm; and the third principle component represents the profitability and financial pressure of a firm. Regarding the leaf node #12-21, all variables are principle component which represents the profitability, liquidity, cash flow ability and financial difficulty of a firm. Regarding the leaf node #12-22, its first principle components consists of two ratios, DR and Z-score, which represent the debt paying ability of a firm; its second principle component consists of one ratio, CFR, which represents the cash flow ability of a firm; its third principle component consists of two ratio, ROA and CFAR, which represents the profitability and the cash flow ability of a firm. Regarding leaf node #12-23, its first principle component consists of three ratios (ROA, DR and Z-score), which represent the profitability, debt paying ability and financial health of a firm; and its second principle component consists of three liquidity related ratios (CR, CFR and CFAR) which represent the liquidity of a firm.

Regarding the leaf node #13-21, its first principle component consists of three liquidity related ratios (CR, QR and CFAR); its second principle component consists of one earning related and one debt related ratios (ROA and DR); and its third principle component consists of one corporate governance related and one financial healthy related ratios (SPR and Z-score). Hence, the first principle component represents liquidity of a firm; the second principle component represents the profitability and debt paying ability of a firm; and the third principle component represents the financial pressure and financial health of a firm. Regarding the leaf node #13-22, it has three principle components. The fist principle component consists of two ratios (QR and DR) which represent the debt paying ability of a firm. The second principle component consists of one ratio, CFAR, which represents the cash flow ability of a firm. The third principle component consists of two ratios (CR and SPR), which represents the liquidity and corporate governance health of a firm. Regarding the leaf node #13-23, it has two principle components. The first principle component consists of four ratios

((ROA, CR, QR, SPR)), which represents the profitability, liquidity and financial pressure of a firm; its second principle component consists of four ratios (DR, CFR, CFAR, Z-score), which represents the debt paying ability, cash flow ability and financial health of a firm. Regarding the leaf node #13-24, its first principle component consists of four ratios (CR, QR, CFR and CFAR), which represents the liquidity of a firm; and its second principle components consists of three ratios (ROA, DR, and Z-score) which represent the profitability, debt paying ability and the financial health of a firm. The third principle component consists of one ratio, SPR, which represents the corporate governance health of a firm.

Regarding the leaf node #14-21, its first principle component consists of two liquidity related ratios and one corporate governance related ratios (CFR, CFAR and SPR); its second principle component consists of the debt related ratios (CR, QR and DR); and its third principle component consists of one financial healthy related ratios (Z-score). Hence, the first principle component represents the liquidity of a firm; the second principle component represents the debt paying ability of a firm; and the third principle component represents the financial health of a firm. Regarding the leaf node #14-22, it has three principle components. The fist principle component consists of three ratios (QR, CFR and CFAR), which represent the liquidity of a firm. The second principle component consists of two ratios (DR and Z-score), which represent the debt paying ability and the financial health of a firm. The third principle component consists of three ratios (ROA, CR, and SPR), which represent the profitability, the liquidity and the corporate governance health of a firm. Regarding the leaf node #14-23, its first principle component consists of four ratios (CR, DR, CFAR and Z-score), which represent the liquidity, debt paying ability and the financial health of a firm; its second principle component consists of two ratios (ROA and QR), which represent the profitability and debt paying ability of a firm; its third principle component consists of one ratios (CFR), which represent the cash flow ability of a firm. Regarding the leaf node #14-24, its first principle component consists of four ratios (ROA, CFR, CFAR and Z-score), which represents the profitability, cash flow ability and financial health of a firm; and its second principle component consists of three ratios (CR, QR and DR), which represent the debt paying ability of a firm.

We can efficiently exploit one single group or compare different groups from comparing the similarity of each extracted features provided by PCA. As Canbas et al.

(2005) had done, an early warning model for the observations can be estimated according to these major factor loadings, such as discriminant, logit, probit, and ANN. By applying PCA to the financial data, the important financial factors can be used to explain the FFR patterns under a certain financial conditions of a firm. In sum, the experimental results show that the proposed quantitative approach with the GHSOM and PCA is helpful in obtaining useful features and can be used to help detect deception regarding FFR or other financial distress scenarios.

4.6 Modeling phase – pattern-extracting module

We take two leaf nodes: #11 and #14-21 as an example to explain about uncovering the regularity of FFR fraud categories from the corresponding indictments and sentences for major securities crimes issued by the Department of Justice. Based on the ten FFR fraud categories discussed in Beasley et al. (1999), Table 17 summarizes the FFR fraud categories commonly adopted by companies clustered in these two leaf nodes. The code and year in the first two columns indicate respectively the company SIC code and the year of financial statements clustered.

Table 17. Common FFR fraud categories within #11 and #14-24.

Code	year	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10
leaf node #11											
2505	1998	●									
2529	1998						●		●		
8716	1999						●		●		
2334	1999						●		●		
3039	2004	●									
1601	1998								●		
1221	2002	●							●		●
1221	2003	●							●		●
2014	2003	●							●		
5901	1997						●		●		
5901	1998						●		●		
5901	1999						●		●		
leaf node #14-24											
2206	1999								●		
2350	1998								●		
2407	2002	●			●	●		●	●		●
2407	2003	●			●	●		●	●		●
2407	2004	●			●	●		●	●		●
2490	2000	●							●		
2490	2002	●							●		
8295	1998				●				●		
1221	2001	●							●		●
8723	1998				●				●	●	
2017	1997				●				●		
5007	1999				●				●		

FC1: recording fictitious revenues; FC2: recording revenues prematurely;

FC3: no description/overstated about revenues; FC4: overstating existing assets;

FC5: recording fictitious assets or assets not owned; FC6: capitalizing items that should be expensed;

FC7: understatement of expenses/liabilities; FC8: misappropriation of assets;

FC9: inappropriate disclosure; FC10: other miscellaneous techniques.

*The code and year in the first two columns indicate the company code and the year of each clustered financial statement.

As shown in Table 17, the common FFR fraud categories found in leaf node #11 are recording fictitious revenues (FC1), capitalizing items that should be expensed (FC6) and misappropriation of assets (FC8). The common FFR fraud categories found in leaf node #14-24 are recording fictitious revenues (FC1), overstating existing assets (FC4) and misappropriation of assets (FC8). With further traces back to the corresponding indictments and sentences, even though both of these two groups have recording fictitious revenues (FC1), we find that the ways of committing this fraud category are quite different.

For instance, some fraud samples (3039 and 1221) in leaf node #11 were found using FC1 via creating fictitious transactions and defrauding export drawbacks from the Internal Revenue Service by reporting fictitious export sales. Moreover, some fraud samples (1601and 1221) used FC8 by processing the receipt and payment in advance. In contrast, some fraud samples (2407, 8723, and 2017) in leaf node #14-24 were found to use FC4 through purchasing intangible asset/long-term investment with high premiums. Some fraud samples (2206, 2407, 2490, 8723, and 2017) used FC8 through related party transactions and merger and acquisition activities to misappropriate cash.

In sum, Table 17 shows that the observed corporate behaviors (i.e., common FFR fraud categories extracted based upon the associated indictments) in different leaf nodes are distinctive even though these nodes are clustered based upon the corporate financial situations proxied by the input variables.

The overall FFR fraud categories extracted from each leaf node of FT are summarized in Table A1. We summarize the common FFR fraud categories into Table 18 for further comparison.

Table 18. Summary of the common FFR fraud categories.

leaf node	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10
#11	●					●		●		
#12-21			●							
#12-22	●			●		●			●	
#12-23		●		●		●		●	●	
#12-24	●								●	
#13-21							●	●		
#13-22						●	●			
#13-23				●			●	●		
#13-24								●	●	
#14-21						●		●	●	
#14-22	●							●		
#14-23						●	●	●		
#14-24	●			●				●		

FC1: recording fictitious revenues; FC2: recording revenues prematurely;
FC3: no description/overstated about revenues; FC4: overstating existing assets;
FC5: recording fictitious assets or assets not owned; FC6: capitalizing items that should be expensed;
FC7: understatement of expenses/liabilities; FC8: misappropriation of assets;
FC9: inappropriate disclosure; FC10: other miscellaneous techniques.

In Table 18, we find those leaf nodes from the same branch tend to have similar common fraud categories. For example, the branch #12 has common fraud categories FC1, FC4, FC6 and FC9. The branch #13 has common fraud categories FC7 and FC8. The branch #14 has common fraud categories FC1, FC6 and FC8. This phenomenon may be resulted from the nature of the SOM, that is, in the topological space of the SOM, the nodes (i.e., groups) with similar features tend to be located nearby. Therefore, the overall distribution of fraud categories in FT can also reveal more information as the FFR knowledge map, which can contribute to build up the knowledge base for FFR detection. Besides, we believe that as the amount of training samples keep accumulated, the represented patterns (i.e., fraud categories) of FT will become more solid and reliable.

4.7 Analyzing phase – group-finding, classifying module

In the analyzing phase, we conclude that the non-fraud-central rule is better through comparing the corresponding sum of (type I and type II) classification errors (Table 12 and Table 14). The dominance of the non-fraud-central rule leads to an implication that most of fraud samples cluster around the non-fraud counterpart.

In analyzing phase, both training samples and testing samples are classified based on the non-fraud-central rule. As shown in Table 19, the testing samples consist of 182 firm-year observations which comprise 54 fraud samples and 128 non-fraud samples over the period from 2002 to 2008. All these testing samples are different from the training samples.

Table 19. The list of fraud and non-fraud firms in testing samples.

No	Fraud firm	SIC code	Fraud year	Non-fraud firm	SIC code	Sampling period
1	雅新	2418	2003-2007	瑞昱	2379	2002-2008
2	遠航	5605	2003-2006	陸海	5603	2002-2008
3	友昱	3506	2004-2006	新能	3196	2002-2008
4	名鐘	6276	2008	元山	6275	2002-2008
5	宏億	3079	2005	華新科	2492	2002-2008
6	亨豐科	6242	2007-2008	迅杰	6343	2002-2008
7	合邦	6103	2004-2007	金麗科	3228	2002-2008
8	東森	2614	2002-2007	寶成	9904	2002-2008
9	歌林	1606	2002-2008	大亞	1609	2002-2008
10	仕欽	6232	2004-2007	佳鼎科	5318	2002-2008
11	勤美	1532	2007-2008	利奇	1517	2002-2008
12	邰港	3350	2004-2008	律勝	3354	2002-2008
13	飛寶動能	4413	2004-2007	國隆	6502	2002-2008
14	新泰伸	5017	2004-2007	榮剛	5009	2002-2008

The other arrangements regarding the data and the significant variable selection are the same as the ones in the training phase.

The classification results of applying the non-fraud-central rule to the training samples and the testing samples are shown in Table 20. The type I error and type II error for both training and testing samples are lower than 20% that implies an acceptable prediction performances regarding the finance application.

Table 20. The classification result.

	type I error	type II error
Training samples	13.62%	13.28%
Testing samples	11.54%	19.78%

In the modeling phase of our proposed decision support approach, different preference of the parameters w_1 (the weight of type I error importance) and w_2 (the weight of type II error importance) for obtaining the parameters β_1 and β_2 can generate different classification boundaries for the fraud-central rule and non-fraud-central rule, respectively. The subjective criteria in selecting a suitable parameter set of w_1 and w_2 consider the issue of the trade-off phenomenon regarding the classification error; that is, defining an acceptable prediction performance of the model in the training stage. We believe that letting the decision makers decide their own acceptable prediction performance based on their domain knowledge or experience in the model training stage can make the proposed decision approach more reliable and useful for a specific application domain, such as the fraud detection issue in this study.

4.8 Decision support phase – feature-retrieving module

The results of the feature-retrieving module come from the pattern-extracting module and the feature-extracting module. We illustrate the results in the following subsections.

4.8.1 Retrieve from pattern-extracting module

Table A2 summarizes the commonly adopted FFR fraud categories of the testing samples identified as the fraud class in all leaf nodes of the FT. The identification

performance of the FFR fraud categories are shown in Table 21, and the detail identification performance of each leaf node is summarized in Table A3.

Table 21. The overall FFR fraud categories identification performance.

true	non-fraud	fraud	fraud	non-fraud	Accuracy 100-average(type I, type II)
predict	fraud	non-fraud	fraud	non-fraud	
error	(type I error)	(type II error)			
percentage	17.53%	13.77%	86.23%	82.47%	84.35%

The overall type I error means based on the fraud categories of a leaf node, what percentage of non-fraud categories are misidentified. The overall type II error means based on the non-fraud categories, what percentage of fraud categories are misidentified. The overall accuracy is the average percentage of the correct fraud categories percentage and the correct non-fraud categories percentage. The overall type I error is 17.53%, type II error is 13.77%, and the accuracy is 84.35%. According to Table 21, the results can effectively support the decision making process for FFR identification.

For further discussion, we take the leaf node #11 and leaf node #14-24 as an example to give a detail description. As shown in Table A2 and Table A3, the common FFR fraud categories found in leaf node #11 are recording fictitious revenues (FC1) and misappropriation of assets (FC8), which fit two of three common FFR fraud categories in #11 retrieved from the modeling phase. The common FFR fraud categories found in leaf node #14-24 are recording fictitious revenues (FC1), overstating existing assets (FC4) and misappropriation of assets (FC8), which fit all of three common FFR fraud categories in #14-24 retrieved from the training stage. In sum, the feature extraction mechanism can actually catch the most common FFR patterns of the testing samples. The experimental results show that the implementation of the DSS architecture based on the proposed dual approach with the feature extraction mechanism is helpful in obtaining FFR features and can be used to help detect FFR. That is, the extracted common FFR fraud categories are integrated with the results of PCA feature extraction to point out the relevant input variables, which can be further associated with the common FFR fraud categories, and provide a clear inference for any risky investigated sample that facilitate the investigation of decision makers.

Such cluster results are derived from the competitive learning nature of the GHSOM, which works as a regularity detector and is supposed to discover statistically salient features of the sample population (Rumelhart and Zipser, 1985). That is, there are no predefined categories into which samples are to be classified; rather, the GHSOM must develop its own feature representation of the sample which captures the most salient features of the population of sample. Furthermore, through a ton of small-sized mappings, the GHSOM classifies the sample into more subgroups with hierarchical relationships instead of a dichotomous result and therefore further and more delicate analyses are feasible.

4.8.2 Retrieve from feature-extracting module

The retrieved principle components for any leaf node can be applied to be linked with its retrieved fraud categories and then provide an inference about its potential fraud behavior. Take the testing samples belonged to the leaf node #11 and #14-24 as examples, the retrieved principle components for the leaf node #11 and #14-24 are shown in Table 22.

Table 22. The principle components retrieved by the feature-retrieving module for the testing samples within #11 and #14-24.

Leaf node	Principle component	Variable	Description
#11	1	CR	debt paying ability
	2	QR, CFR, CFAR	liquidity
	3	ROA, SPR	profitability, financial pressure
#14-24	1	ROA, CFR, CFAR, Z-score	profitability, cash flow ability, financial health
	2	CR, QR, DR	debt paying ability

For the leaf node #11, the first principle component represents the debt paying ability of a firm. The second principle component represents the liquidity of a firm. The third principle component represents the profitability and financial pressure of a firm. Regarding the leaf node #14-24, its first principle component represents the

profitability, cash flow ability and financial health of a firm. The second principle component represents the debt paying ability of a firm.

Then, the retrieved principle components are linked to its common fraud categories of the testing sample, which can be used to explain the rationality of the provided principle components. For example, the common FFR fraud categories found in leaf node #11 are recording fictitious revenues (FC1), capitalizing items that should be expensed (FC6) and misappropriation of assets (FC8), which may be caused from lacking of debt paying ability, liquidity, profitability, or under server financial pressure. The common FFR fraud categories found in leaf node #14-24 are recording fictitious revenues (FC1), overstating existing assets (FC4) and misappropriation of assets (FC8), which may be caused from bad cash flow ability or weak debt paying ability. Therefore, for any unknown investigated sample classified into #11 and #14-24, the results with both fraud categories and principle components also help provide the possible clues as the direction for further inspection.

4.9 Analyzing phase – decision-supporting module

Based on the common FFR fraud categories observed in the leaf node, we further investigate the causes of the observed common FFR fraud categories with the assistance of experts with domain knowledge to identify the relevant input variables of such regularity for future financial reporting. That is, the identified common FFR fraud categories of each leaf node are further integrated with the principal components extracted from the classified samples. Such information can help identify the relevant input variables as the pre-warning signal, which reveals the potential fraudulent activities, for any samples clustered into this investigated leaf node by the GHSOM.

Without losing the generalization, the results of the decision-supporting module are shown in Table 23. The information contains both features and patterns could provide clues to facilitate decision making. The explanation and speculation are mainly done by the decision makers.

Let's take the investigated samples predicted to commit fraud and belonged to the leaf nodes #11 or #14-24 to descript part of the results of the decision-supporting module. For any investigated sample identified fraud and belonged to the leaf node #11

or #14-24, the associated investigation could be summarized as follows for decision aid. Any sample belonged to leaf node #11 may have a liquidity pressure and a weakness of short term debt paying ability such that they tend to commit FFR through recording fictitious revenues, capitalizing items that should be expensed, or misappropriating assets approach. For the leaf node #14-24, any sample belonged to it may have a bad cash flow condition and worse profitability, thus the overall financial pressure such that they tend to commit FFR through overstating existing assets, or recording fictitious revenues approach.

Table 23. The results of decision-supporting module for any investigated sample identified fraud.

Leaf node	FFR fraud categories	Principle components	Description
#11	FC1, FC6, FC8	{CR, (QR, CFR, CFAR), (ROA, SPR)}	{debt paying ability, liquidity, (profitability, financial pressure)}
#12-21	FC3	{(ROA, CR, QR, DR, CFR, CFAR, Z-score)}	{(profitability, liquidity, cash flow ability, financial difficulty)}
#12-22	FC1, FC4, FC6, FC9	{(DR, Z-score), CFR, (ROA, CFAR)}	{debt paying ability, cash flow ability, (profitability, the cash flow ability)}
#12-23	FC2, FC4, FC6, FC8, FC9	{(ROA, DR, Z-score),(CR, CFR, CFAR)}	{(profitability, debt paying ability, financial health), liquidity}
#13-21	FC7, FC8	{(CR, QR, CFAR), (ROA, DR), (SPR, Z-score)}	{liquidity, (profitability, debt paying ability), (financial pressure, financial health)}
#13-22	FC6, FC7	{(QR, DR, Z-score), (CFR, CFAR), (CR, SPR)}	{debt paying ability, cash flow ability, (liquidity, corporate governance health)}
#13-23	FC4, FC7, FC8	{(ROA, CR, QR, SPR), (DR, CFR, CFAR, Z-score)}	{(profitability, liquidity and financial pressure), (debt paying ability, cash flow ability, financial health)}

#13-24	FC8, FC9	{(CR, QR, CFR, CFAR), (ROA, DR, Z-score), SPR}	{liquidity, (profitability, debt paying ability, financial health), corporate governance health}
#14-21	FC6, FC8, FC9	{(CFR, CFAR, SPR), (CR, QR, DR), Z-score}	{liquidity, debt paying ability, financial health}
#14-22	FC1, FC8	{(QR, CFR, CFAR), (DR, Z-score), (ROA, CR, SPR)}	{liquidity, (debt paying ability, the financial health), (profitability, the liquidity, corporate governance health)}
#14-23	FC6, FC7, FC8	{(CR, DR, CFAR, Z-score), (ROA, QR), CFR}	{(liquidity, debt paying ability, financial health), (profitability, debt paying ability), cash flow ability}
#14-24	FC1, FC4, FC8	{(ROA, CFR, CFAR, Z-score}, (CR, QR, DR)}	{(profitability, cash flow ability, financial health), debt paying ability}

FC1: recording fictitious revenues; FC2: recording revenues prematurely;
FC3: no description/overstated about revenues; FC4: overstating existing assets;
FC5: recording fictitious assets or assets not owned; FC6: capitalizing items that should be expensed;
FC7: understatement of expenses/liabilities; FC8: misappropriation of assets;
FC9: inappropriate disclosure; FC10: other miscellaneous techniques.

Based on our analysis, many fraud samples belonged to Iron & Steel and Building Material & Construction industries and committed FFR in 1998 and 1999 during East Asian Financial Crisis. The operation of fraud firms deteriorated sharply due to the bear market and could not generate sufficient net cash inflow. They committed FFR to conceal the embezzlement and other undesirable outcomes from investors and creditors. Prevalent FFR fraud categories include overstating revenues through fictitious sales, embezzling money via accounts such as temporary payment or prepayment for purchases, recording loans from related party into accounts receivable. These FFR behaviors make some accounts falsified, such as accounts receivable, related party transaction or other relevant input variables.

5. Methods comparison

We also compare the results of our proposed method with the ones of three existing methods — the Support Vector Machine (SVM) (Vapnik, 1995), the SOM with Linear Discriminant Analysis (LDA) (named SOM+LDA) (Carlos, 1996), the (traditional) GHSOM with LDA (named GHSOM+LDA), SOM, BPNN and Decision Tree (DT).[4] The SVM is a supervised learning method which has specialty in recognizing patterns, and has been widely used for classification and regression analysis (Vapnik, 1995; Hsu et al., 2009). The idea of applying the SOM with LDA in FFD is derived from Carlos's (1996) study. In contrast, the GHSOM+LDA method uses all training samples to construct merely one GHSOM tree while the clustering module of our proposed dual approach uses the fraud and non-fraud samples to construct FT and NFT, respectively. The trial of GHSOM+LDA is used to justify the effectiveness of the analyzing phase of the proposed dual approach.

5.1 SVM

Given a training set of instance-label pairs (x_i, y_i), $i = 1,\ldots,l$ where $x_i \in R^n$ and $y \in \{1,-1\}^l$, the SVM (Boser et al., 1992; Vapnik, 1995; Cortes and Vpaink, 1995; Hsu et al., 2010) require the solution of the following optimization problem:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i \\ &s.t. \quad y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ &\xi_i \geq 0. \end{aligned} \tag{13}$$

Here training vectors x_i are mapped into a higher dimensional space by the function ϕ. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. C > 0 is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function. There are four basic kernels:

[4] We use the GHSOM toolbox in the platform of Matlab R2007a, the SVM and the SOM package in the platform of SPSS Clementine 12.0, and the LDA package in the platform of SPSS.

Linear: $K(x_i, x_j) = x_i^T x_j$.

Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$.

Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0$.

Sigmoid: $K(x_i, x_j) = \text{tsaih}(\gamma x_i^T x_j + r)$.

Here, γ, r and d are kernel parameters. The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis. Therefore, we choose RBF as the kernel function of SVM.

5.2 SOM+LDA

As stated in Carlos's (1996) study, the SOM+LDA which, on the basis of the information contained in a multidimensional space — in the case exposed, financial ratios — generates a space of lesser dimensions. In this way, similar input patterns are represented close to one another on a map. Such neural networks can be combined with other mathematical models applied to the prediction of corporate failure. From among all these, without doubt the most popular is LDA. For example, Canbas et al. (2005) proposed a methodological framework for constructing the integrated early warning system (IEWS) that can be used as a decision support tool in bank examination and supervision process for detection of banks, which are experiencing serious problems. Well known multivariate statistical technique (principal component analysis), was used to explore the basic financial characteristics of the banks, and discriminant, logit and probit models were estimated based on these characteristics to construct IEWS.

Based on the idea of below studies, we use SOM to cluster the training samples, and then apply LDA in each node of SOM as a classifier to identify the fraud samples. Table 24 describes the habitual working procedure of SOM+LDA.

Table 24. The habitual working procedure of the SOM+LDA.

step 1:	Sample and measure variable.
step 2:	Identify the significant variables that will be used as the input variables.
step 3:	Use the training samples to set up an SOM.
step 4:	For each node of SOM, set up a LDA model.
step 5:	For each investigated sample *s*, identify the winning node *x* of SOM.
step 6:	Use the trained LDA of the node *x* to predict the investigated sample *s*.

The map size of the SOM is 4 × 3, as shown in Figure 11.

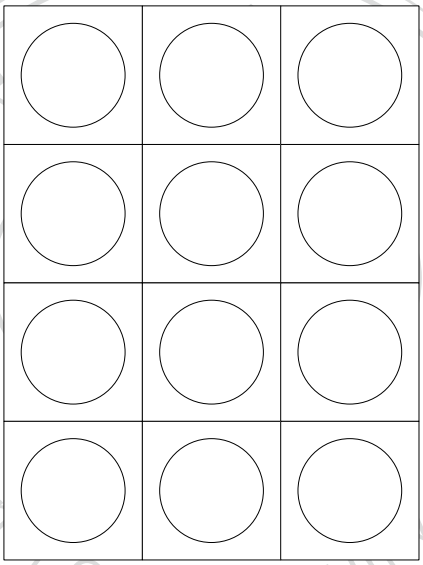


Figure 11. The map size of the SOM in the SOM+LDA method.

5.3 GHSOM+LDA

The trial of GHSOM+LDA method is used to justify the effectiveness of the analyzing phase of the proposed dual approach, which uses all training samples to construct merely one GHSOM tree while the training phase of our proposed dual approach uses the fraud and non-fraud samples to construct FT and NFT, respectively.

The GHSOM is used to cluster the training samples, and then apply LDA in each leaf node of the GHSOM as a classifier to identify the fraud samples. Table 25 describes the habitual working procedure of GSOM+LDA.

Table 25. The habitual working procedure of the GHSOM+LDA.

step 1:	Sample and measure variable.
step 2:	Identify the significant variables that will be used as the input variables.
step 3:	Use the training samples to set up a GHSOM.
step 4:	For each node of GHSOM, set up a LDA model.
step 5:	For each investigated sample *c*, identify the winning node *x* of GHSOM.
step 6:	Use the trained LDA of the node *x* to predict the investigated sample *c*.

The parameter $\tau_1 = 0.8$ and $\tau_2 = 0.07$ are set for the GHSOM+LDA method. The GHSOM+LDA method results in a GHSOM tree with 16 leaf nodes, named one tree (ONET), which is shown Figure 12.

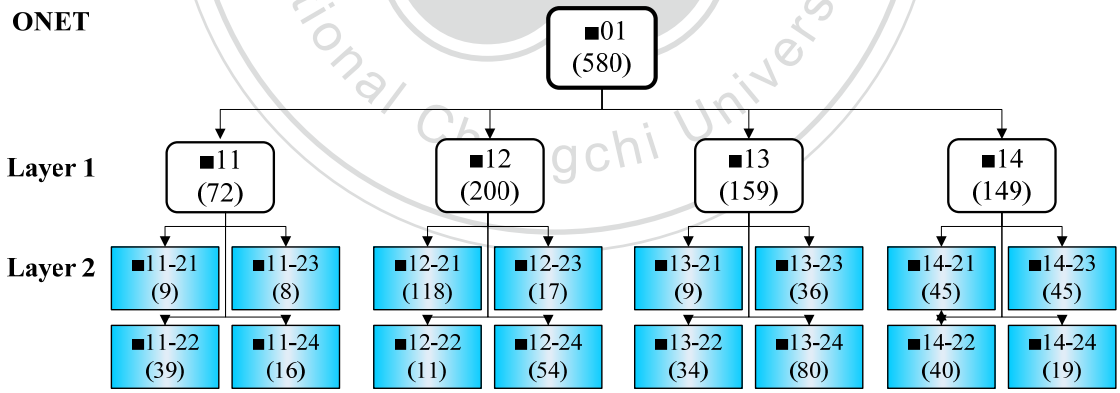


Figure 12. The obtained GHSOM tree of the GHOM+LDA method.

5.4 SOM

Based on the theoretical feature of SOM, samples with similar internal feature tend to be grouped together. Following this principle, one out sample can use the feature of its belonging node of SOM as its represented feature. The nodes of SOM with high risk are marked in the training stage, and any sample classified into any of these high risk nodes shell be considered high risk too based on the nature of unsupervised learning.

We use SOM to cluster the training samples, and then determined the risk level in each node of SOM. The node with high fraud sample proportion will be classified as a fraud group. Any sample classified into this group will be predicted as a fraud one. Table 26 describes the working procedure of the SOM method.

Table 26. The habitual working procedure of the SOM.

step 1:	Sample and measure variable.
step 2:	Identify the significant variables that will be used as the input variables.
step 3:	Use the training samples to set up an SOM.
step 4:	For each node of SOM, determined the risky node with high FFR proportion.
step 5:	For each investigated sample *c*, identify the winning node *x* of SOM.
step 6:	If sample *c* is classified into one of the risky nodes of SOM, it will be predicted as a fraud one, otherwise, a non-fraud one.

The map size of the SOM is 4 × 3 is the same as the SOM+LDA method which is shown in Figure 13. Here we set the fraud sample proportion bigger than 13% as the risky nodes. The determined risky nodes resulted from step 4 is shown in Figure 13, in which the risky nodes are marked in color.

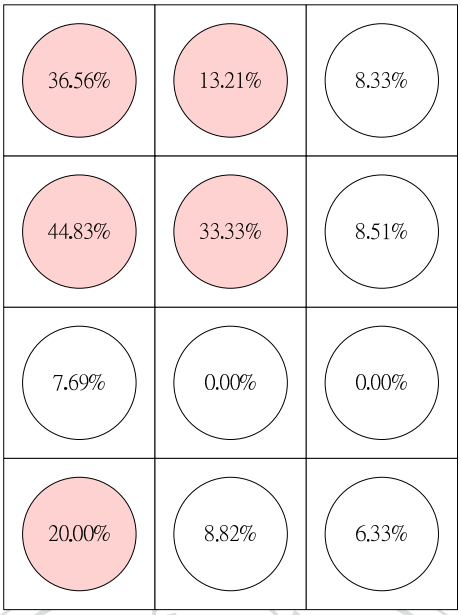


Figure 13. The map size of the SOM with FFR proportions.

The risky nodes of the obtained SOM also reveal a certain regularity of SOM that similar nodes tend to be located nearby. The comparative location of nodes can also help decision makers get the background of an investigated sample from the nodes nearby its belonging node. The boundary of risky and healthy nodes can be easily found through the information of FFR sample proportion, other indicator for help observe the boundary are not also worth of trying as a part of classifier.

This study also compares the prediction performance regarding another two data mining tools, back-propagation neural network (BPNN) and decision tree (DT), to compare their prediction performance and to discuss how the results of each method could contribute to discover FFR.

5.5 BPNN

The Back propagation neural network (BPNN) is proposed by Rumelhart et al. (1986). The BPNN is a supervised learning neural network tool, which is one of the popular techniques for classification and prediction. The training of BPNN by steepest descent method (SDM) is to follow negative gradient direction of cost function to find out the optimal weighting and bias.

In this study, the BPNN structure is shown in Figure 14 and the parameter setting of BPNN is set as follows: the transfer function is sigmoid, the iterations are set to 10000, the learning rate is set to 0.01, and the momentum factor is set to 0.9. The BPNN is run 10 times with different initial weights, and then calculate the minimal, average and maximal prediction error. The minimal prediction error is chosen as the represented performance of BPNN. The weights of BPNN are shown in Table 27. The classification results of BPNN are shown in Table 28.

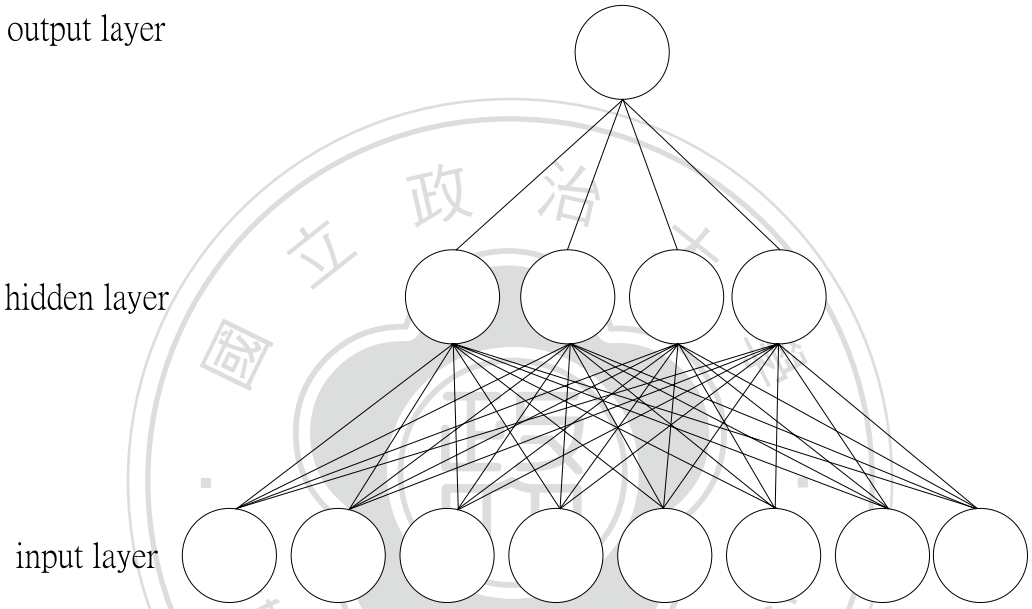


Figure 14. The BPNN structure.

Table 27. The weights of BPNN.

variable	weight
CFR	1.996
DR	0.1973
Z-score	0.1865
CR	0.1093
SPR	0.0902
CFAR	0.0837
QR	0.0731
ROA	0.0604

Table 28. The classification results of the BPNN.

No.	training stage			testing stage	
	Type I error	Type II error	Sum of errors	Type I error	Type II error
1	15.2%	32.74%	47.95%	10.94%	64.81%
2	13.7%	44.25%	57.95%		
3	19.27%	34.51%	53.79%		
4	20.56%	33.63%	54.19%		
5	20.13%	32.74%	52.87%		
6	16.49%	40.71%	57.20%		
7	17.99%	35.40%	53.39%		
8	17.13%	39.82%	56.95%		
9	15.42%	42.48%	57.90%		
10	20.77%	31.86%	52.63%		
	minimal prediction error:		47.95%		
	average prediction error:		57.95%		
	maximal prediction error:		54.48%		
represent	15.2%	32.74%	47.95%	10.94%	64.81%

5.6 DT

Decision Tree (DT) is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

ID3 (Iterative Dichotomiser 3) was developed in 1986 by Ross Quinlan. The algorithm creates a multi-way tree, finding for each node (i.e., in a greedy manner) the categorical feature that will yield the largest information gain for categorical targets. Trees are grown to their maximum size and then a pruning step is usually applied to improve the ability of the tree to generalize to unseen data.

C4.5 (Quinlan, 1993) is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e., the output of the ID3 algorithm) into sets of if-then rules. The accuracy of each rule is then evaluated to determine the order in

which they should be applied. Pruning is done by removing a rule's precondition if the accuracy of the rule improves without it. C5.0 (Quinlan, 1996) uses less memory and builds smaller rule sets than C4.5 while being more accurate.

In this study, we use C5.0 algorithm to build up the DT. The minimum number per child branch is 10. The obtained DT structure and the obtained rule sets are shown in Figure 15 and Figure 16.

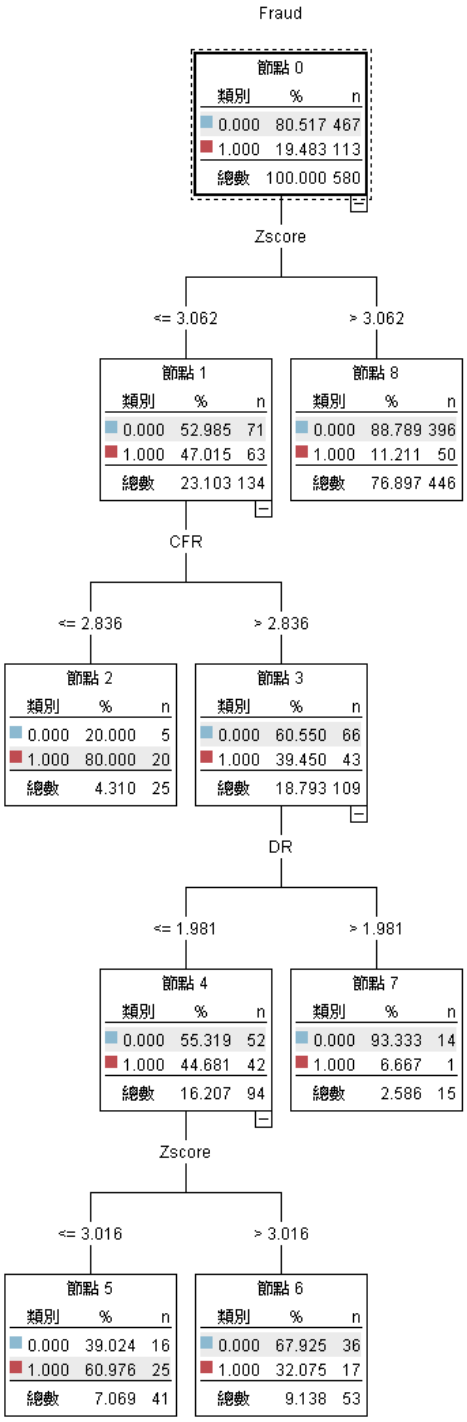


Figure 15. The obtained DT structure.

```
Zscore <= 3.062 [ Mode: 0 ]
 CFR <= 2.836 [ Mode: 1 ] => 1.0
 CFR > 2.836 [ Mode: 0 ]
      DR <= 1.981 [ Mode: 0 ]
           Zscore <= 3.016 [ Mode: 1 ] => 1.0
           Zscore > 3.016 [ Mode: 0 ] => 0.0
      DR > 1.981 [ Mode: 0 ] => 0.0
Zscore > 3.062 [ Mode: 0 ] => 0.0
```

Figure 16. The obtained DT rules.

The experimental designs for both training stage and testing stage are the same as the previously mentioned settings, and the results are shown in Table 29.

Table 29. The experimental results of our dual approach, the SVM, SOM+LDA, GHSOM+LDA, SOM, BPNN and DT methods.

	Training samples		Testing samples	
	type I error	type II error	type I error	type II error
dual approach	13.62%	13.28%	11.54%	19.78%
SVM	24.84%	21.24%	45.31%	27.78%
SOM+LDA	16.49%	30.09%	32.81%	51.85%
GHSOM+LDA	20.34%	22.12%	22.66%	44.44%
SOM	34.48%	35.4%	26.56%	57.41%
BPNN	15.2%	32.74%	10.94%	64.81%
DT	4.5%	60.18%	15.63%	68.52%

Compared with the SVM, SOM+LDA, GHSOM+LDA, SOM, BPNN, and DT methods, the corresponding sum of (type I and type II) classification errors of the analyzing phase in the proposed dual approach is much lower in most cases. As shown in Table 29, the type I error and the type II error for both training samples and testing

samples of our approach are lower than 20% that implies acceptable classification performance regarding the FFD application.

5.7 Discussion of the experimental results

This research proposes a DSS architecture based on a novel dual approach that includes the training phase, the modeling phase, the analyzing phase and the decision support phase. In the training phase, the GHSOM generates more subgroups instead of dichotomous outcomes which can facilitate delicate analysis. In the modeling phase, the classification rule is built and the fraud related features of the samples of a leaf node is extracted for evaluating any sample classified into this leaf node. The analyzing phase applies the dominant classification rule obtained from the modeling phase. The decision support phase integrated the classification result and the associated features for decision aid. The results confirm the feasibility of the proposed dual approach which can contribute to FFR detection.

Specifically, the GHSOM generates more subgroups instead of dichotomy and provides more delicate features embedded in the samples. Additionally, the unsupervised learning nature of the GHSOM renders a more robust clustering compared to traditional dichotomous classification. Each sample of the GHSOM is treated equally without specifying its fraud attribute, and the clustering results can help the following steps of rule forming and feature extracting.

The methods comparison results show that the implementation of the DSS architecture based on the proposed dual approach not only outperforms the supervised methods such as the SVM, BPNN, DT and the unsupervised methods such as the SOM+LDA and GHSOM+LDA, but also reveals more domain information such as the extracted features (principal components) and the extracted patterns (fraud categories) based on the common characteristics of the same group (leaf node), and the spatial relationship among fraud and non-fraud samples. We believe that the improvement of the GHSOM can contribute to the applicability in FFD, and can provide an alternative way of data mining which enriches the background knowledge retrieved from the similar samples of the historical data, and this is the potentiality of our proposed dual approach.

6. Discussions and implications

This study develops the dual approach as a DSS architecture that can be used for FFD. The proposed dual approach is data-driven to perform the system modeling via directly using the sampled data. As shown in Figure 5, the system architecture based on the dual approach consists of a series of four phases. The details and the associated modules have been explained phase by phase in section 3.1 to section 3.4, respectively. Below, we summarize these four phases and the corresponding modules.

In the training phase, the data preprocessing is first executed through the sampling module and variable-selecting module. Then, all samples with the corresponding values of selected variables are the input of the clustering module to generate two GHSOMs (i.e., fraud samples are used to generate FT and non-fraud samples are used to generate NFT). The modeling phase consists of the statistic-gathering module, rule-forming module, feature-extracting module and pattern-extracting module. The first two modules utilize the statistics of FT and NFT leaf nodes to form the classification rules which are different due to different spatial hypotheses. Then, the classification rules are tuned respectively and compete with each other to become the dominant one. The last two modules involve the discovery of features (e.g., principal components) and patterns (e.g., fraud categories) in the FT leaf nodes. The extracted features and patterns of each FT leaf node are valuable for FFD decision support through being retrieved in the decision support phase.

The analyzing phase consists of the group-finding module and classifying module. Based on the GHSOM clustering rule, each investigated sample is clustered into its belonging leaf nodes in FT and NFT, and these two leaf nodes are paired. Then, the classifying module uses the dominant classification rule obtained from the training phase to determine if the investigated sample is fraud. If an investigated sample is identified fraud, then the decision support phase will be executed. The feature-retrieving module retrieves the features and patterns from the investigated sample's belonging FT leaf node, and the decision-supporting module integrates the extracted features and patterns for the purpose of decision aid.

The implications for decision support in FFD, the research implications, and the FFR managerial implications are given in the following subsections.

6.1 The decision support in FFD

The proposed system architecture results in a process of identifying any interesting pattern that can facilitate the FFD decision making. Besides, the dual approach can be integrated with other statistical, mathematical, artificial intelligence, or machine learning techniques to extract and identify useful information which contribute to the domain knowledge.

Ngai et al. (2010) have done a complete academic review of FFD. They summarized that the data mining techniques of outlier detection and visualization have seen only limited use. In real world FFD cases, the sample size of the fraud cases compared with the normal majority is relatively low. The detection of the fraud case may be regarded as recognizing the outlier from the healthy majority. Therefore, Agyemang et al. (2006) pointed out that outlier detection is a very complex task akin to find a needle in a haystack. Although we use the pair-matching to do the sampling in the FFR case mentioned in Chapter 4, as shown in section 3.1, the proposed sampling module does not stick on the pair-matching. Since the dual approach is data-driven, it can be applied to the case of outlier detection in FFD.

With the implementation of the proposed dual approach based on the GHSOM, the fraud samples and non-fraud samples are clustered separately and then the matched pairs of groups can help scale down the focus scope, such that the developed classification rule based on the associated spatial hypotheses (i.e., non-fraud-central or fraud-central) is capable of identifying the fraud samples (i.e., outliers) more accurately. Note that the classification rule based on a spatial hypothesis is developed through the proposed optimization technique for the corresponding discriminant boundary, in which the decision makers can objectively set their weightings of type I and type II errors. Therefore, the dominate classification rule is flexible enough when applying to other FFD application domains with different preference of type I and type II errors.

Also, providing fraud related patterns for a suspected sample can contribute to FFD decision making. The feature-extracting module and pattern-extracting module is able to be applied to other financial fraud scenarios (e.g., bank fraud, insurance fraud) and financial crises scenarios (e.g., bankruptcy, stock market crashes). When applying

to other similar scenarios, the feature-extracting module changes the input variables according to the problem domain, and the pattern-extracting module adjusts the definition of fraud categories (or crisis categories) to develop the pattern map of FT. Such reference can enhance the quality of decision support by pinpointing the risk area (i.e., the variables in the principal components, and the fraud categories) required attention, and therefore help reduce the likelihood of issuing doubtful loan-related decisions and help provide sufficient information for decision support.

It is worthy of a future work to implement the proposed DSS architecture based on the dual approach for any FFD related application domain. The implemented DSS may contain an additional data-importing module, and a visualization module. The visualization module visualizes the identification results for a creation investigated sample, and provides a whole viewpoint of the FT (i.e., pattern map) in which the fraud categories and the principal components of each leaf node can be selected to be shown on the diagram. The decision support module can be extended to include the results of other feature extracting mechanisms (such as statistical approach and data mining approach). Then, a voting mechanism will be used to integrate all the obtained features to help decision makers receive equitable and rational decision support.

6.2 The research implications

This study utilizes the advantage of the GHSOM and pioneers a novel dual approach for constructing a DSS architecture for FFD purpose. The proposed DSS architecture is data-driven and adaptive to fit any FFD scenarios with two basic groups, fraud and non-fraud (unhealthy and healthy), and the fraud group can be divided into different subcategories which represent distinctive fraud patterns. The designed modules and processes are described and evaluated phase by phase, and the methods within several modules (sampling, variable-selecting, clustering, feature-extracting, and pattern-extracting modules) can be replaced with other similar methods which make the proposed DSS architecture more generalizable for the real world practical use.

The experimental results show that the implementation of the DSS architecture based on the proposed dual approach can help the decision support in FFD through

providing an alternative way of investigating financial data, which includes the dual clustering by the GHSOM and the development of adaptive classifiers for each pairs of subgroups (i.e., leaf nodes).

The implementation of the proposed DSS architecture can not only identify the fraud cases, but also provide the extracted features and patterns for reference. Furthermore, the clustering results in FT can provide more amounts of subgroups, and provide more fraud-related information within subgroups compared to the dichotomous detection results which are generally provided by the conventional FFD studies, so that a comprehensive exploration of the relationship between different subgroups is intriguing and possible. Also, the GHSOM of the proposed system architecture is applicable to the adaptive sample size (i.e., data-driven) since the GHSOM will be re-developed accordingly, and the feature-extracting module and pattern-extracting module can provide the corresponding characteristics (e.g., the inherent variable features and the fraud patterns) as the fraud potentiality for the investigated samples.

Different from the traditional GHSOM studies which cluster the whole training samples at one time, the clustering module of the proposed approach separates the training samples into fraud group and non-fraud group to generate two GHSOMs. The idea of such design is to improve the unsupervised learning mechanism through utilizing the spatial relationship between a pairs of leaf nodes from these two GHSOMs. That is, for each pair of leaf nodes, developing an adaptive classification rule based on such spatial relationship. The discriminate boundary can be tuned through the proposed optimization method in which the weightings of type I and type II errors are adjustable according to the decision makers' preference that renders the outcome of the analyzing phase with more acceptable classification performance for a certain application domain.

For each leaf node of FT, the feature extraction mechanism extracts the fraud categories from the exogenous information and the principal components from the input variables, respectively. Therefore, for any sample clustered into a leaf node of FT, the corresponding principal components and fraud categories can be used to represent the associated fraud regularities. These fraud regularities can be used as the pre-warning signal and can reveal the associated potential fraud activities to help monitor the suspected sample. Furthermore, the pattern-extracting module needs a definition of domain categories from some authentic references. The pattern-extracting

module can be implemented through either the domain experts or applying the text mining technique.

The theoretical meaning of the spatial relationship is an interesting topic and is worthy of a deeper analysis. The spatial hypotheses (or belief) of this study are: for a pair of leaf nodes from FT and NFT, the associated fraud samples tend to locate around the non-fraud counterparts, or the associated non-fraud samples tend to locate around the fraud counterparts. In the modeling phase of the dual approach, the spatial relationships between the fraud samples and the non-fraud samples of the paired subgroups are identified and then utilized to develop the associated classification rule which is the dominant classification rule of two candidate classification rules (non-fraud-central rule and fraud-central rule) derived from two spatial hypotheses. The dominance of the non-fraud-central rule leads to an implication that most of fraud samples cluster around the non-fraud counterpart, and the dominance of fraud-central rule leads to an implication that most of non-fraud samples cluster around the fraud counterpart. If one of these two spatial regularities fits to the sample data, the corresponding classification rule can provide superior classification performance; moreover, the spatial relationship within fraud and non-fraud samples can provide valuable insights for the FFD domain experts.

The above mentioned implications bring out the advantages of the outcome of the proposed dual approach. That is, the abundant information associated with the outcome could enrich the conventional dichotomous detection for decision aid.

6.3 The FFR managerial implications

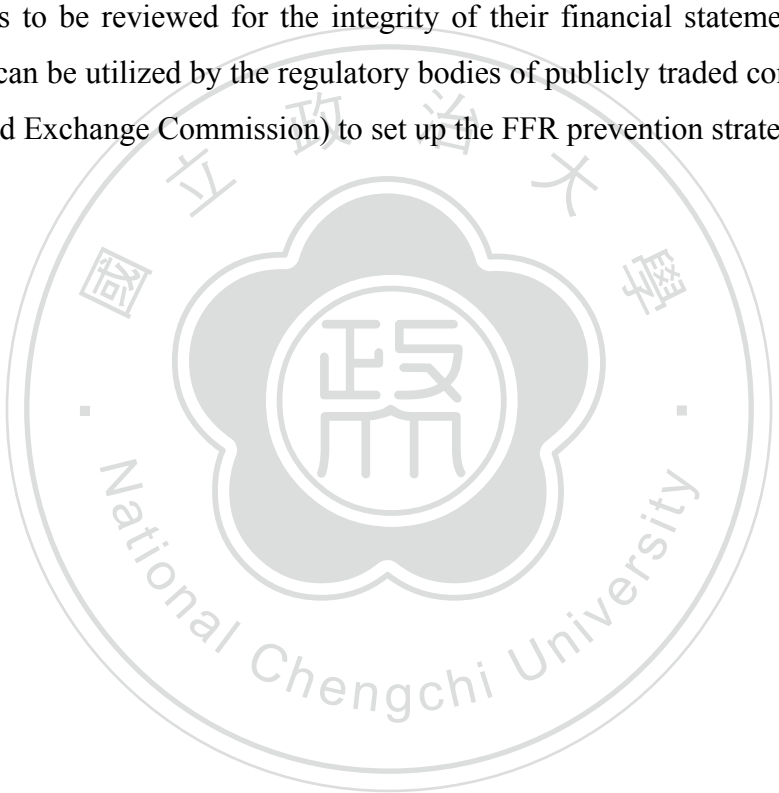
In contrast with prior FFR studies focusing on finding the signification input variables regarding FFR and providing dichotomous prediction result without giving further explanations, this study has shown that the proposed dual approach can help not only identify FFR, but also help interpret the FFR behaviors of samples.

The proposed approach involves a feature inspection on the fraud training samples, and the accumulated FFR understandings help creditors and capital providers evaluate the integrity of financial statements to facilitate their investment or credit decision-making. The accumulated FFR understandings also help facilitate the

development of credit risk evaluation model used internally. Besides, the feature results can be employed by auditors into their audit plans to ensure their firms or clients remain competitive.

Regarding the implication for forensic accounting, the retrieved information including the FFR fraud categories and the principal components can help forensic accountants by providing the common features based on the similar samples belonged to the same leaf node, and help them perform extended procedures as part of the statutory audit.

Furthermore, the clustering results of the proposed approach can help give the list of companies to be reviewed for the integrity of their financial statements, and such information can be utilized by the regulatory bodies of publicly traded companies (e.g., Securities and Exchange Commission) to set up the FFR prevention strategies.



7. Conclusion

Because of the nature of competitive learning, the GHSOM, an unsupervised neural networks extended from the SOM, can work as a regularity detector that is supposed to help discover statistically salient features of the sample population (Hogan et al., 2008). With the spatial correspondent hypotheses, this study presents a DSS architecture with four phases based on the proposed dual approach for FFD decision support, in which two GHSOMs (i.e., fraud samples are used to generate FT and non-fraud samples are used to generate NFT) are generated in the training phase. In the modeling phase, for each leaf node of FT, a feature extraction mechanism including the feature-extracting module and pattern-extracting module is developed to provide the associated fraud related features, and the extracted features will be used as a part of the evaluation for any risky investigated sample. The classification rules are formed to help identify fraud cases through applying the adaptive classification rules into each pair of fraud and non-fraud subgroups from FT and NFT. In the analyzing phase, the dominant classification rule is applied to examine the investigated samples. For the investigated samples which have been identified fraud, the relevant fraud categories and variables are retrieved and integrated in the decision support phase. All the provided information is helpful for the decision making process of FFD.

Unlike the traditional approach applying the SOM in FFD (Carlos, 1996) which uses all training samples to generate one SOM, our proposed DSS architecture takes advantage of being able to generate two GHSOMs (FT and NFT), in each of which two spatial hypotheses — for each pair of leaf nodes from FT and NFT, the fraud (or non-fraud) samples are cluster around their counterparts— are set to create the candidate classification rules. That is, using the statistic information among samples from different GHSOMs helps respectively generate the non-fraud-central and fraud-central rules. These two rules are tuned via inputting all samples to determine the optimal discrimination boundary of each candidate classification rule within each pair of leaf nodes from NFT and FT. This study derives the optimization technique that renders adjustable and effective rules for classifying fraud and non-fraud samples. The decision makers can objectively set their weightings of type I and type II errors. The candidate classification rule that dominates another is adopted as the classification rule in the following analyzing phase. The dominance of the non-fraud-central rule leads to

an implication that most of fraud samples cluster around the non-fraud counterpart, meanwhile the dominance of fraud-central rule leads to an implication that most of non-fraud samples cluster around the fraud counterpart.

To the best of our knowledge, this is the first work that employs the GHSOM to provide topological insights of high-dimensional inputs in addition to hierarchical features. It is worth noting that the implementation of the DSS architecture based on the proposed dual approach is beyond the traditional unsupervised learning approach for FFD through developing a more delicate classifier that can reveal the spatial relationship among fraud and non-fraud subgroups, and the proposed feature extraction mechanism provides more information to represent the potential fraud behaviors for any suspected investigated sample, as a result, support the practical FFD decision making process.

Our preliminary result on FFR experiment confirms the spatial relationship among fraud and non-fraud financial statements, and has better classification performance than the SVM, SOM+LDA, GHSOM+LDA, SOM, BPNN and DT methods. Therefore, for cases with the regularity of the proposed two topological relationships among fraud and non-fraud samples, the implemented DSS architecture based on the proposed dual approach can perform well; furthermore, compared with conventional methods for FFD, the feature extracting results also add more fraud-related characteristics for the investigated samples which are identified fraud.

The limitations of this study would be: (1) compared with other FFD scenarios, the sample size for the FFR issue is limited, (2) subjective parameter setting of the GHSOM, (3) the fraud patterns are various depend on the focused FFD scenario and the results of the pattern-extracting module need to be verified by the domain experts, and (4) the proposed DSS architecture does not evaluated or refined practically until a system prototype is actually being developed.

Future works are suggested as follows: (1) derive the theoretical justification of the rule-forming module in the modeling phase, (2) improve the discrimination boundary setting in the rule-forming module with more sensitivity via an enhanced optimization approach for developing the classification rules, or try other good classifiers, (3) use other clustering methods in the clustering module and compare the results of classifying module in terms of the classification performance and the dominate classification rule derived from which spatial hypothesis, (4) improve the

pattern-extracting module with systematic tools, and (5) conduct experiments on other FFD applications.

Reference

Aas, K., and Eikvil, L. (1999). *Text categorization: A survey*, Technical Report, 941, Norwegian Computing Center.

Agyemang, M., Barker, K., and Alhajj, R. (2006). A comprehensive survey of numeric and symbolic outlier mining techniques, *Intelligent Data Analysis*, 10(6), 521–538.

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance*, 23(4), 589–609.

American Institute of Certified Public Accountants (AICPA) (2002). Statement on Auditing Standards No. 99: Consideration of Fraud in a Financial Statement Audit [Electronic Version]. http://www.aicpa.org/download/members/div/auditstd/AU-00316.PDF.

Antonio, S.A., David, M.G., Emilio, S.O. Alberto, P., Rafael, M.B., and Antonio, S. L. (2008). Web mining based on Growing Hierarchical Self-Organizing Maps: Analysis of a real citizen web portal, *Expert Systems with Applications*, 34(4), 2988–2994.

Association of Certified Fraud Examiners (ACFE) (1998). *1998 Report to the nation on occupational fraud and abuse*, ACFE, Austin, TX.

Association of Certified Fraud Examiners (ACFE). (2008). *2008 Report to the nation on occupational fraud and abuse*, ACFE, Austin, TX.

Basens, B., Setiono, B., Mues, C., and Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation, *Management Science*, (49:3), 312–319.

Beasley, M.S., Carcello, J.V., and Hermanson, D.R. (1999). *Fraudulent financial reporting: 1987–1997: An analysis of U.S. public companies*, COSO, New York.

Bell, T. B., and Carcello, J. V. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting, *Auditing: A Journal of Practice & Theor*, (19:1), 169–184.

Bond, C. F., and DePaulo, B. M. (2006). Accuracy of deception judgments, *Personality and Social Psychology Review*, 10(3), 214–234.

Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers, *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ACM Press, 144–152.

Budayan, C. (2008). Strategic group analysis: Strategic perspective, differentiation and performance in construction, Doctoral dissertation, Middle East Technical University.

Budayan, C., Dikmen, I., and Birgonul, M. T. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping, *Expert Systems with Applications*, 36(9), 11772–11781.

Canbas, S., Cabuk, A., and Kilic, S.B. (2005). Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case, *European Journal of Operational Research*, 166, 528–546.

Carlos, S. C. (1996). Self organizing neural networks for financial diagnosis, *Decision Support System*, 17, 227–2386.

Claessens, S., Djankov, S., and Lang, L. H. P. (2000). The separation of ownership and control in East Asian Corporations, *Journal of Financial Economics*, 58(1-2), 81–112.

Cortes, C., and Vapnik, V. (1995). Support-vector network, *Machine Learning*, 20, 273–297.

Daskalaki, S., Kopanas, I., Goudara, M., and Avouris, N. (2003). Data mining for decision support on customer insolvency in telecommunications business, *European Journal of Operational Research*, 145, 239–255.

Dechow, P. M., Ge, W., Larson, C. R., Sloan, R. G., and Investors, B. G. (2007). Predicting material accounting manipulations, *AAA 2007 Financial Accounting and Reporting Section (FARS) [Electronic Version]*. http://ssrn.com/abstract=997483.

Dechow, P.M., Sloan, R.G., and Sweeney, A.P. (1996). Cause and consequences of earnings manipulation: an analysis of firms subject to enforcement actions by the SEC," *Contemporary Accounting Research*, 13(1), 1–36.

Desai, V. S., Crook, J. N., and Overstreet, J. (1996). A comparison of neural networks and linear scoring models in the credit union environment, *European Journal of Operational Research*, 95, 24–37.

Dittenbach, M., Merkl, D., and Rauber, A. (2000). The Growing hierarchical self-organizing map, *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks- IJCNN 2000*.

Dittenbach, M., Rauber, A., and Merkl, D. (2002). Uncovering hierarchical structure in data using the growing hierarchical self-organizing map, *Neurocomputing*, 48(1-4), 199–216.

Eklund, T. (2002). Assessing the feasibility of self organizing maps for data mining financial information, *ECIS 2002*, Gdansk, Poland.

Elliot, R., and Willingham, J. (1980). *Management fraud: detection and deterrence*, Petrocelli, New York, NY.

Fanning, K.M., and Cogger, K.O. (1998). Neural network detection of management fraud using published financial data, *International Journal of Intelligent Systems in Accounting, Finance & Management*, 7(1), 21–41.

Farber, D. B. (2005). Restoring trust after fraud: does corporate governance matter?, *Accounting Review*, 80(2), 539–561.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd Edition, Academic Press, London.

Granzow, M., Berrar, D., Dubitzky, W., Schuster, A., Azuaje, F. J., and Eils, R. (2001). Tumor classification by gene expression profiling: Comparison and validation of five clustering methods, *SIGBIO Newsletter*, 21(1), 16–22.

Green, P., and Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology, *Auditing: A Journal of Practice & Theory*, 16(1), 14–28.

Guo, Y., Hu, J., and Peng, Y. (2011). Research on CBR system based on data mining, *Applied Soft Computing*, 11(8), 5006–5014.

Hoogs, B., Kiehl, T., Lacomb, C., and Senturk, D. (2007). A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud, *Intelligent Systems in Accounting Finance and Management*, (15:1/2), 41–56.

Hsu, K. Y. (2008). Exploring financial reporting fraud, M.A. Thesis, National Chengchi University, Department of Management Information System.

Hsu, S. H., Hsieh, J. P. A., Chih, T. C., and Hsu, K. C. (2009). A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression, *Expert Systems with Applications*, 36(4), 7947–7951.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2010). A practical guide to support vector classification, Technical Report, National Taiwan University.

Huang, S. Y., Tsaih, R. H., and Lin, W. Y. (2012). Unsupervised Neural Networks Approach for Understanding Fraudulent Financial Reporting, *Industrial Management & Data Systems*, 112(2), 224–244.

Huang, S. Y., and Tsaih, R. H. (2012). The Prediction Approach with Growing Hierarchical Self-Organizing Map, *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 838-844.

Huang, S. Y., and Tsaih, R. H., Fang, Y. (2012). The Dual Approach for Decision Making, *The 2012 Decision Sciences Institute Annual Meeting (DSI)*, San Francisco, USA.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., and Felix, W. F. (2010). Identification of fraudulent financial statements using linguistic credibility analysis, *Decision Support Systems*, 50(3), 585–594.

Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3), 264–323.

Jiang, J. (1999). Image compression with neural networks - a survey, *Signal Process Image Communication*, 14(9-7), 737–760.

Jolliffe, I. T. (1986). *Principal Component Analysis*, Springer, New York.

Jolliffe, I. T. (2002). *Principal Component Analysis*, second edition, New York: Springer-Verlag New York, Inc.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis, *Educational and Psychological Measurement*, 20, 141–151.

Khan A.U., Sharma, T. K., and Sharma, S. (2009). Classification of Stocks Using Self Organizing Map, *International Journal of Soft Computing Application*s, 4, 19–24.

Klein M, and Methlie, L. B. (1995). *Knowledge-Based Decision Support Systems with Applications in Business*, 2nd edn, Wiley.

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

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, (43), 59–69.

Kohonen, T. (1989), *Self Organization and Associative Memory*, 3rd ed. Springer, Berlin.

Kohonen, T. (1995), *Self-Organizing Maps*, Springer, Berlin.

KPMG Peat Marwick (1998), *Fraud Survey*, KPMG peat Marwick, Montvale, NJ.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., and Vishny, R. (1999). Corporate ownership around the world, *Journal of Finance*, 54(2), 471–517.

Lee, G., T. K. Sung, and Chang, N. (1999). Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction, *Journal of Management Information Systems*, 16, 63–85.

Lee, T. S., Chiu, C. C., Chou, Y. C., and Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, *Computational Statistics & Data Analysis*, 50, 1113–1130.

Lee, T. S., and Yeh, Y. H. (2004). Corporate governance and financial distress: evidence from Taiwan, *Corporate Governance: An International Review*, 12(3), 378–388.

Li, S. (2000). The development of a hybrid intelligent system for developing marketing strategy, *Expert Systems with Applications*, 27, 395–409.

Liu, Y., Yeh, R. H. ,and He, R. (2006). Sea surface temperature patterns on the West Florida Shelf using the Growing Hierarchical Self-Organizing Maps, J. Atmos. *Oceanic Technology*, 23(2), 325–338.

Loebbecke, J. K., Eining, M. M., and Willingham, J. J. (1989). Auditors' experience with material irregularities: frequency, nature, and detectability, *Auditing*, 9(1), 1–28.

Lu, C. J., and Wang, Y. W. (2010). Combining independent component analysis and growing hierarchical self-organizing maps with support vector regression in product demand forecasting, *International Journal of Production Economics*, 128(2), 603–613.

Mangiameli, P., Chen, S. K., and West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods, *European Journal of Operational Research*, 93(2), 402–417.

Maron, M. E. (1961). Automatic indexing: An experimental inquiry, *Journal of the ACM*, 8, 404–417.

Martín-Guerrero, J.D., Palomares, A., Balaguer-Ballester, E., Soria-Olivas, E., Gómez- Sanchis, J., and Soriano-Asensi, A. (2006). Studying the feasibility of a recommender in a citizen web portal based on user modeling and clustering algorithms, *Expert Systems with Applications*, 30(2), 299–312.

Matsatsinis, N. F., Doumpos, M., and Zopounidis, C. (1997), Knowledge acquisition and representation for expert systems in the field of financial analysis, *Expert Systems with Applications*, 12, pp. 247–262.

Mohanty, R. P., and Deshmukh, S. G. (1997). Evolution of an expert system for human resource planning in a petroleum company, *Production Economics*, 51, 251–261.

Newman, D. P., Patterson, E. and Smith, R. (2001). The influence of potentially fraudulent reports on audit risk assessment and planning. *Auditing: A Journal of Practice & Theory*, 76(1), 59–80.

Ngai, E. W, Yong, H. U., Wong, Y. H., Chen, Y., and Sun, X. (2010). 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.

Nguyen M. N., Shi, D., and Quek, C. (2008). A nature inspired Ying-Yang approach for intelligent decision support in bank solvency analysis, *Expert Systems with Applications*, 34(4), 2576–2587

Min, J. H., and Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, *Expert Systems with Applications*, 28, 603–614.

Oja, E. (1992). Principal components, minor components, and linear neural networks, *Neural Networks*, 5(6), 927–935.

Pagano, R. R. (2001). Understanding Statistics in the Behavioral Sciences, Sixth ed., Wadsworth/Thomson Learning, California.

Quinlan, J. R. (1986), Induction of Decision Trees, *Machine Learning*, 1, 81–106.

Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

Quinlan, J. R. (1996), Bagging, boosting, and C4.5., *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, pp. 725-730.

Rasmussen, E. (1992), *Information retrieval: data structures and algorithms*. Prentice-Hall, Inc.

Pearson, K. (1901), On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 2(6), 559–572.

Persons, O. S. (1995), Using financial statement data to identify factors associated with fraudulent financial reporting, *Journal of Applied Business Research*, 11(3), 38–46.

Pinson, S. (1992), A Multi-Expert Architecture for Credit Risk Assessment: The CREDEX system, IN: D.E. O'Lerary and P.R. Watkins (Eds), *Expert Systems in finance*, 37–64.

Rauber, Merkl, A., D., and Dittenbach, M. (2002). The Growing hierarchical self-organizing map: exploratory analysis of high-dimensional data, *IEEE Transactions on Neural Networks*, 13(6), 1331–1341.

Richardson, A. J., Risien, C., and Shillington, F. A. (2003). Using self organizing maps to identify patterns in satellite imagery, *Prog. Oceanogr*, 59, 223–239.

Ringnér, M. (2008), What is principal component analysis?, *Nature Biotechnology*, 26, 303–304.

Risien, C. M., Reason, C. J. C., Shillington, F. A., and Chelton, D. B. (2004). Variability in satellite winds over the Benguela upwelling system during 1999– 2000, J. *Geophys. Res.*, 109, C03010, doi:10.1029/2003JC001880.

Rumelhart, D. E., and Zipser, D. (1985). Feature discovery by competitive learning, *Cognitive Science*, 9, 75–112.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors, *Nature*, 323, 533–536.

Schweighofer, E., Rauber, A., and Dittenbach, M. (2001). Automatic text representation classification and labeling in European law, *International Conference on Artificial Intelligence and Law (ICAIL)*. ACM Press.

Serrano, C. (1996). Self Organizing Neural Networks for Financial Diagnosis, *Decision Support Systems*, 17, 227–238.

Shaw, P. J. A. (2003). *Multivariate statistics for the Environmental Sciences*, Hodder-Arnold.

Shih, J. Y., Chang, Y. J., and Chen, W. H. (2008). Using GHSOM to construct legal maps for Taiwan's securities and futures markets, *Expert Systems with Applications*, 34(2), 850–858.

Shin, H. W., and Sohn, S. Y. (2004). Segmentation of stock trading customers according to potential value, *Expert Systems with Applications*, 27(1), 27–33.

Shin, K. S., Lee, T. S., and Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model, *Expert Systems with Applications*, 28, 127–135.

Shin, K. S., and Lee, Y. J. (2002). A genetic algorithm application in bankruptcy prediction modeling, *Expert Systems with Applications*, 23, 321–328.

Stice, J. D. (1991). Using financial and market information to identify pre-engagement factors associated with lawsuits against auditors, *Accounting Review*, 66(3), 516–533.

Summers, S. L., and Sweeney, J. T. (1998). Fraudulently misstated financial statements and insider trading: An empirical analysis, *Accounting Review*, 73(1), 131–146.

Tabachnick, B. G., and Fidell, L. S. (2001). *Using multivariate statistics*, 4th Edition, Boston: Allyn & Bacon.

Tsai, C. F. (2009). Feature selection in bankruptcy prediction, *Knowledge-Based Systems*, 22(2), 120–127.

Tsaih, R. H., Lin, W. Y., and Huang, S. Y. (2009), Exploring Fraudulent Financial Reporting with GHSOM, *Pacific Asia Workshop on Intelligence and Security Informatics (PAISI)*, *Lecture Notes in Computer Science*, 5477, 31–41.

Turban, E. (1993), *Decision Support and Expert Systems: Management Support Systems*, 3rd eds, Macmillan.

Turk, M. A., and Pentland, A. P. (1991). Eigenfaces for recognition, *Journal of Cognitive Neuroscience*, 3(1), 71–86.

Vapnik, V. N. (1995), *The nature of statistical learning theory*, Springer.

Virdhagriswaran, S., and Dakin, G. (2006). Camouflaged fraud detection in domains with complex relationships, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 941–947.

Weisenborn, D., and Norris, D. (1997). Red flags of management fraud, *National Public Accountant*, 42(2), 29–33.

Wen, W., Wang, W. K., and Wang, T. H. (2005). A hybrid knowledge-based decision support system for enterprise mergers and acquisitions, *Expert Systems with Applications*, 28, 569–582.

Wen, W., Chen, Y. H., and Pao, H. H. (2008). A mobile knowledge management decision support system for automatically conducting an electronic business, *Knowledge-Based Systems*, 21(7), 540–550.

Yeh, Y., T. Lee, and Woidtke, T. (2001). Family control and corporate governance: Evidence from Taiwan, *International Review of Finance*, 2(1-2), 21–48.

Zhang, J., and Dai, D. (2009). An adaptive spatial clustering method for automatic brain MR image segmentation, *Progress in Natural Science*, 19(10), 1373–1382.

Zopounidis, C., Doumpos, M., and Matsatsinis, N. F. (1997). On the use of knowledge-based decision support systems in financial management: a survey, *Decision Support Systems*, 20, 259–277.

Appendix

The overall FFR fraud categories extracted from each leaf node of FT are summarized in Table A1. The common FFR fraud categories within each leaf node are marked with * in the column.

Table A1. Common FFR fraud categories within all leaf nodes of FT.

leaf node #11		*FC1	FC2	FC3	FC4	FC5	*FC6	FC7	*FC8	FC9	FC10
(code)	(year)										
2505	1998	●									
2529	1998						●		●		
8716	1999						●		●		
2334	1999						●		●		
3039	2004	●									
1601	1998								●		
1221	2002	●							●		●
1221	2003	●							●		●
2014	2003	●							●		
5901	1997						●		●		
5901	1998						●		●		
5901	1999						●		●		
leaf node #12-21		FC1	FC2	*FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10
5385	2001				●						●
8713	1999			●							
1918	1998			●						●	
leaf node #12-22		*FC1	FC2	FC3	*FC4	FC5	*FC6	FC7	FC8	*FC9	FC10
2398	2001	●		●	●	●			●	●	
2398	1999	●		●	●	●			●	●	
2494	2002	●									
3001	2000						●			●	
3001	2001						●			●	
3001	1999						●			●	
5385	2000				●						●
6145	2003	●									
6145	2004	●									
6250	2004					●	●			●	

1602	1994				●		●		●		●
leaf node #12-23		FC1	*FC2	FC3	*FC4	FC5	*FC6	FC7	*FC8	*FC9	FC10
8702	1994		●						●		
8702	1995		●						●		
5504	2000		●								
8710	1999		●								
8719	1997						●				
8701	1995			●	●						
8701	1996			●	●						
1602	1995				●		●		●		●
1918	1996			●						●	
1918	1997			●						●	
2101	1997								●	●	●
2613	1999						●			●	
2913	1996				●				●		
leaf node #12-24		*FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	*FC9	FC10
4113	2004	●								●	
leaf node #13-21		FC1	FC2	FC3	FC4	FC5	FC6	*FC7	*FC8	FC9	FC10
8712	1998							●	●		
5503	2000							●		●	●
8719	1998						●				
2014	2001	●							●		
2014	2002	●							●		
8717	1998							●		●	
leaf node #13-22		FC1	FC2	FC3	FC4	FC5	*FC6	*FC7	FC8	FC9	FC10
2553	1999						●	●			
8716	1998						●		●		
8188	2001							●			
8724	2000							●		●	
8724	1999							●		●	
2005	1999								●		
2019	2000						●				
2019	1999						●				
8711	1999						●	●			●
leaf node #13-23		FC1	FC2	FC3	*FC4	FC5	FC6	*FC7	*FC8	FC9	FC10
8705	1998				●	●	●	●	●		
8714	1999			●				●			
8382	1998				●		●		●		

8706	1998				●			●	●		
leaf node #13-24		FC1	FC2	FC3	FC4	FC5	FC6	FC7	*FC8	*FC9	FC10
1505	1998									●	
1505	1999									●	
8708	1998							●	●		
8708	1999							●	●		
2101	1998								●	●	●
2101	1999								●	●	●
leaf node #14-21		FC1	FC2	FC3	FC4	FC5	*FC6	FC7	*FC8	*FC9	FC10
5504	1999		●								
2328	1998		●	●						●	
2334	1998						●		●		
1505	1997									●	
5007	1998				●				●		
2614	1999	●					●		●	●	●
1466	1998			●			●			●	
leaf node #14-22		*FC1	FC2	FC3	FC4	FC5	FC6	FC7	*FC8	FC9	FC10
2505	1997	●									
2328	1997		●	●						●	
2334	1997						●		●		
2350	1997								●		
2398	2000	●		●	●	●			●	●	
2490	2001	●							●		
1601	1997								●		
1602	1996				●		●		●		●
leaf node #14-23		FC1	FC2	FC3	FC4	FC5	*FC6	*FC7	*FC8	FC9	FC10
2206	2000								●		
1436	1997							●			
1436	1998							●			
2553	1998						●	●			
1207	2000	●					●		●	●	●
1207	1998	●					●		●	●	●
1207	1999	●					●		●	●	●
2005	1998								●		
2016	1997				●			●			
2016	1998				●			●			
2017	1996				●				●		
2017	1998				●				●		

2019	1998						●				
8705	1997				●	●	●	●	●		
8708	1997							●	●		
8714	1997			●				●			
8714	1998			●				●			
9911	1998			●						●	
9801	2000	●					●		●	●	●
9801	1998	●					●		●	●	●
9801	1999	●					●		●	●	●
leaf node #14-24		*FC1	FC2	FC3	*FC4	FC5	FC6	FC7	*FC8	FC9	FC10
2206	1999								●		
2350	1998								●		
2407	2002	●			●	●		●	●		●
2407	2003	●			●	●		●	●		●
2407	2004	●			●	●		●	●		●
2490	2000	●							●		
2490	2002	●							●		
8295	1998				●				●		
1221	2001	●							●		●
8723	1998				●				●	●	
2017	1997				●				●		
5007	1999				●				●		

Note: The common FFR fraud categories of each leaf node are marked with *.

Table A2 summarizes the commonly adopted FFR fraud categories of the testing samples identified as the fraud class in all leaf nodes of the FT. The code and year in the first two columns indicate the company SIC code and the year of financial statements. The common FFR fraud categories extracted from the feature-extracting module are marked in gray. The common FFR fraud categories within each leaf node are marked with * in the column.

Table A2. Common FFR fraud categories of the testing samples.

leaf node #11		*FC1	FC2	FC3	FC4	FC5	FC6	FC7	*FC8	FC9	FC10
1606	2008			●					●		
2418	2006	●							●		
2418	2008	●							●		
4413	2005	●							●		
leaf node #12-22		*FC1	FC2	FC3	*FC4	FC5	FC6	FC7	FC8	FC9	FC10
3506	2003	●	●		●		●				
3506	2004	●	●		●		●				
6232	2006	●									
6232	2007	●									
6103	2008	●			●				●		
3079	2005				●					●	
5017	2003									●	
1606	2006			●					●		
leaf node #12-23		FC1	*FC2	FC3	FC4	FC5	*FC6	FC7	*FC8	FC9	FC10
3506	2002	●	●		●		●				
1532	2008				●		●		●	●	
5605	2002		●				●		●		
5605	2003		●				●		●		
5605	2004		●				●		●		
5605	2005		●				●		●		
leaf node #14-22		*FC1	FC2	FC3	FC4	FC5	FC6	FC7	*FC8	FC9	FC10
6232	2002	●									
6103	2004	●			●				●		
3350	2004								●		
2418	2004	●							●		
leaf node #14-24		*FC1	FC2	FC3	*FC4	FC5	FC6	FC7	*FC8	FC9	*FC10
6103	2005	●			●		●		●	●	
6103	2006	●			●		●		●	●	
2614	2008	●			●				●		●
2614	2006	●			●				●		●
2614	2007	●			●				●		●

FC1: recording fictitious revenues;
FC2: recording revenues prematurely;
FC3: no description/overstated about revenues;
FC4: overstating existing assets;
FC5: recording fictitious assets or assets not owned;
FC6: capitalizing items that should be expensed;
FC7: understatement of expenses/liabilities;
FC8: misappropriation of assets;
FC9: inappropriate disclosure;
FC10: other miscellaneous techniques.

The identification performance of the FFR fraud categories are summarized in Table A3, in which the classification errors (type I error and type II error) are calculated.

Table A3. The identification performance of the FFR fraud categories.

leaf node		true	predict	fraud-> fraud	fraud-> non-fraud	true	predict	non-fraud-> non-fraud	non-fraud ->fraud
#11									
1606 (code)	2008 (year)	1,8	1,6,8	100.00%	0.00%	2,3,4,5,6,7,9,10	2,3,4,5,7,9,10	87.50%	12.50%
2418	2006	1,8	1,6,8	100.00%	0.00%	2,3,4,5,6,7,9,10	2,3,4,5,7,9,10	87.50%	12.50%
2418	2008	1,8	1,6,8	100.00%	0.00%	2,3,4,5,6,7,9,10	2,3,4,5,7,9,10	87.50%	12.50%
4413	2005	3,8	1,6,8	50.00%	50.00%	1,2,4,5,6,7,9,10	2,3,4,5,7,9,10	75.00%	25.00%
#12-22		true	predict	fraud-> fraud	fraud-> non-fraud	true	predict	non-fraud-> non-fraud	non-fraud ->fraud
3506	2003	1,2,4,6	1,4,6,9	75.00%	25.00%	3,5,7,8,9,10	2,3,5,7,8,10	83.33%	16.67%
3506	2004	1,2,4,6	1,4,6,9	75.00%	25.00%	3,5,7,8,9,10	2,3,5,7,8,10	83.33%	16.67%
6232	2006	1	1,4,6,9	100.00%	0.00%	2,3,4,5,6,7,8,9,10	2,3,5,7,8,10	66.67%	33.33%
6232	2007	1	1,4,6,9	100.00%	0.00%	2,3,4,5,6,7,8,9,10	2,3,5,7,8,10	66.67%	33.33%
6103	2008	1,4,8	1,4,6,9	66.67%	33.33%	2,3,5,6,7,9,10	2,3,5,7,8,10	71.43%	28.57%
3079	2005	4,9	1,4,6,9	100.00%	0.00%	1,2,3,5,6,7,8,10	2,3,5,7,8,10	75.00%	25.00%
5017	2003	9	1,4,6,9	100.00%	0.00%	1,2,3,4,5,6,7,8,10	2,3,5,7,8,10	66.67%	33.33%
1606	2006	3,8	1,4,6,9	0.00%	100.00%	1,2,4,5,6,7,9,10	2,3,5,7,8,10	62.50%	37.50%
#12-23		true	predict	fraud-> fraud	fraud-> non-fraud	true	predict	non-fraud-> non-fraud	non-fraud ->fraud

3506	2002	1,2,4,6	2,4,6,8,9	75.00%	25.00%	3,5,7,8,9,10	1,3,5,7,10	66.67%	33.33%
1532	2008	4,6,8,9	2,4,6,8,9	100.00%	0.00%	1,2,3,5,7,10	1,3,5,7,10	83.33%	16.67%
5605	2002	2,6,8	2,4,6,8,9	100.00%	0.00%	1,3,4,5,7,9,10	1,3,5,7,10	71.43%	28.57%
5605	2003	2,6,8	2,4,6,8,9	100.00%	0.00%	1,3,4,5,7,9,10	1,3,5,7,10	71.43%	28.57%
5605	2004	2,6,8	2,4,6,8,9	100.00%	0.00%	1,3,4,5,7,9,10	1,3,5,7,10	71.43%	28.57%
5605	2005	2,6,8	2,4,6,8,9	100.00%	0.00%	1,3,4,5,7,9,10	1,3,5,7,10	71.43%	28.57%
#14-22		true	predict	fraud->fraud	fraud->non-fraud	true	predict	non-fraud->non-fraud	non-fraud->fraud
6232	2002	1	1,8	100.00%	0.00%	2,3,4,5,6,7,8,9,10	2,3,4,5,6,7,9,10	88.89%	11.11%
6103	2004	1,4,8	1,8	66.67%	33.33%	2,3,5,6,7,9,10	2,3,4,5,6,7,9,10	100.00%	0.00%
3350	2004	8	1,8	100.00%	0.00%	1,2,3,4,5,6,7,9,10	2,3,4,5,6,7,9,10	88.89%	11.11%
2418	2004	1,8	1,8	100.00%	0.00%	2,3,4,5,6,7,9,10	2,3,4,5,6,7,9,10	100.00%	0.00%
#14-24		true	predict	fraud->fraud	fraud->non-fraud	true	predict	non-fraud->non-fraud	non-fraud->fraud
6103	2005	1,4,6,8,9	1,4,8	60.00%	40.00%	2,3,5,7,10	2,3,5,6,7,9,10	100.00%	0.00%
6103	2006	1,4,6,8,9	1,4,8	60.00%	40.00%	2,3,5,7,10	2,3,5,6,7,9,10	100.00%	0.00%
2614	2008	1,4,8	1,4,8	100.00%	0.00%	2,3,5,6,7,9,10	2,3,5,6,7,9,10	100.00%	0.00%
2614	2006	1,4,8	1,4,8	100.00%	0.00%	2,3,5,6,7,9,10	2,3,5,6,7,9,10	100.00%	0.00%
2614	2007	1,4,8	1,4,8	100.00%	0.00%	2,3,5,6,7,9,10	2,3,5,6,7,9,10	100.00%	0.00%
Average				86.23%	13.77% (type II)			82.47%	17.53% (type I)

According to Table A3, the identification performance regarding the FFR fraud category is quite well. For each FT leaf node, the fraud categories extracted by the pattern-extracting module can cover most of the common fraud categories of the testing samples (see the ‘fraud->fraud’ column) and the can cover most of the common excluded fraud categories (see the ‘non-fraud->non-fraud’ column). Overall, the results can effectively support the decision making process for FFR identification.