CHAPTER 4
COMPARATIVE ANALYSIS OF EXTENSIVELY USED DATA MINING TECHNIQUES FOR DETECTION OF FINANCIAL STATEMENT FRAUD

4.1 Introduction
4.2 Extensively Used Data Mining Techniques
   4.2.1 Neural Networks
   4.2.2 Decision Tree
   4.2.3 Genetic Programming
   4.2.4 Bayesian Belief Network
4.3 Comparative Analysis of Data Mining Techniques
4.4 Effectiveness of Data Mining Techniques
4.5 Summary

4.1 INTRODUCTION
Financial statement fraud (FSF) costs billions of dollars every year to the world’s economy. Generally, the perpetrators of fraud lie within the organisations. In more than 40% cases fraud has been perpetrated by top executives including board members, directors etc. (Figure 4.1) [KPMG08].

Detecting management fraud is essential once prevention mechanism has failed. Primarily, auditors are supposed to detect falsified financial statement, but in most of the cases auditors are deceived by managers. The estimation of likelihood of fraudulent financial reporting is based on presence or absence of number of fraud risk factors.

Effective discriminators of fraudulent financial reporting include the following factors [Crowder97].
   a. Rapid growth
   b. Weak control environment
   c. Management overly preoccupied with meeting earnings projections
   d. Management that lied to the auditors or that was overly evasive
   e. Ownership status
f. An aggressive management attitude towards financial reporting

![Figure 4.1: Different Perpetrators of Fraud (KPMG India Fraud Survey Report (Forensic) 2008)](image)

Considering these factors along with the application of data mining techniques on financial statements, organisations may be classified as fraud and non-fraud organisations. Several data mining algorithms have been implemented for successful identification of fraudulent financial reporting.

The review of the academic literature suggests the four commonly used data mining techniques namely, Neural Networks, Decision Trees, Genetic Algorithms and Bayesian Belief Networks for detection of financial statement fraud. This chapter investigates the effectiveness of the four techniques in the identification of fraudulent financial statements. In addition, the four techniques are compared in terms of their performances based on eight varying parameters. Neural network appeared as most extensively used technique for detection and identification of financial statement fraud. The rest of the chapter discusses applicability of four extensively used data mining techniques for detection of financial statement fraud (Section 4.2) followed by Section 4.3 which present a comparative analysis of the data mining techniques on the basis of eight parameters. Section 4.4 discusses the effectiveness of four commonly used data mining techniques on the basis of comparative analysis. Section 4.5 summarizes the chapter.
CHAPTER - 4  COMPARATIVE ANALYSIS OF EXTENSIVELY USED DATA MINING TECHNIQUES

4.2 EXTENSIVELY USED DATA MINING TECHNIQUES

Data mining plays an important role in detection of financial statement fraud, as it is often applied to extract and uncover the hidden knowledge, unknown patterns behind very large quantities of data. [Turban07] define data mining as a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequently gain knowledge from a large database. This capability of data mining techniques has been extensively used for detection financial statement fraud. Four data mining techniques are discussed below:

4.2.1 NEURAL NETWORK

A neural network has been described as a “type of artificial intelligence” which uses case based reasoning and pattern recognition to simulate the way the human brain processes and stores information. The key element of this paradigm is composed of a large number of highly interconnected processing elements called neurons, working in unanimity to solve specific problems. These neurons are distributed in a few hierarchical layers and generally contain three types of layers: input, hidden, and output. After receiving the input from all the neurons from an input layer, the values are added through applied weights and converted to an output value by applying an activation function. Then, the result is passed to all of the neurons in the next layer, which provide a feed forward path to the output layer. An iterative training process is applied to adjust the weights between two neurons in two adjacent layers while training samples are presented to the network.

Neural networks are capable of learning the characteristics of potentially fraudulent financial statements by comparing new information to stored data and detecting hidden patterns with in large data set. After leaning the pattern of input data from sample fraud and non – fraud cases, neural network can evaluate the individual data signals to create a distinct behaviour pattern which classify input data as fraudulent or non – fraudulent. The resultant pattern is then applied to detect the presence of fraud in financial statements.

4.2.2 DECISION TREES

A decision tree is a logical model represented as a binary tree usually constructed using a training data set. A Decision Tree helps in predicting the value of a target variable by using a
set of predictor variables. It consists of hierarchically organized sets of rules. Decision tree is a simple recursive structure for representing a decision procedure in which a new instance is classified into one of the predefined classes.

Decision tree attempts to divide observations in mutually exclusive subgroups. Each node in a decision tree corresponds to a set of record from the original data set. The topmost node is named as “root” node and represents all of the rows in the given dataset.

The nodes which have their child are known as “interior” nodes and represent a test on an attribute. The remaining nodes are known as “terminal” or “leaf” nodes and denote a decision class. Each branch of a decision tree represents an outcome of the test.

A process of recursive partitioning is used in construction of a decision tree by dividing the rows in a node into two child nodes. The selection of attribute that best separates the sample is a major concern in construction of a decision tree. Each node is further divided into child nodes until either the subgroups are too small to undergo similar meaningful division or no statistically significant subgroups are produced by splitting further. This successive division of sample may results in a large tree and some of the branches may reflect anomalies in form of outliers or false values. Such branches are required to be removed. This process of removing splitting nodes is known as tree pruning. Tree pruning should be performed in a manner that does not affect the models accuracy rate significantly. The attribute values of the object are tested against the splitting nodes of the decision tree for successful classification of a previously unseen object. According to this test, a path is traced that will conclude with the object’s class prediction. The main advantages of decision trees are that they provide a meaningful way of representing acquired knowledge and make it easy to extract IF-THEN classification rules [Kirkos07].

4.2.3 GENETIC ALGORITHM

Genetics algorithm follows the evolutionary fundamentals of natural science and genetics. It is an adaptive heuristic search algorithm. Genetic algorithm states that the genetic pool of a specific population probably contains the solution for a given problem. Each and every solution to the given problem is represented as a chromosome or genome.
CHAPTER - 4  COMPARATIVE ANALYSIS OF EXTENSIVELY USED DATA MINING TECHNIQUES

The first step of genetic algorithm is random generation of a set of solutions termed as population. This algorithm believes that new population will be better than the old one. This idea progresses the algorithm by creating a new population from the existing one. Genetic operators such as selection, initialization, mutation and crossover are applied on the population of solutions for obtaining an evolved solution in order to find the best possible solution. Fitness of each individual in the population is calculated. New or better solutions are further evolved from already existing solutions on the basis of their fitness. This process of evolution continues until certain condition such as, no further improvement of the best solution is possible, is satisfied.

One of the most important applications of genetic algorithm is detection of fraud. This capability of genetic algorithm has been used by [Hoogs07] for finding falsified financial statements. The genetic algorithm approach takes advantage of expanded information including comparative views of financial metrics and ratios, and the relationships between these comparative metrics over time. The comparative metrics capture current company performance within the context of historical and industry performance. The patterns produced by genetic algorithm comprise combinations of the comparative metrics across multiple fiscal periods, thus capturing multi-quarter interactions of context-driven performance metrics. The algorithm selects pattern variables from a set of eighty five comparative metrics and company characteristics, covering a wide range of financial health indicators. Because the patterns consider multiple fiscal periods, there are multiple opportunities to detect indicators of fraud, making the patterns robust to occasional missing values in relevant metrics. Combinations of patterns in which each pattern captures the same type of behaviour as the other patterns, but uses different metrics, can also mitigate the impact of metrics that have missing values for specific subsets of the population.

4.2.4 BAYESIAN BELIEF NETWORKS

Bayesian belief networks are also known as “belief networks”, “causal probabilistic networks”, “causal nets”, and “graphical probability networks”. Bayesian Belief Networks (BBN) allow for the representation of dependencies among subsets of attributes. A BBN is a directed acyclic graph, where each node represents an attribute and each arrow represents a probabilistic dependence. If an arrow is drawn from node A to node B, then A is parent of B
and B is a descendent of A. In a Belief Network each variable is conditional independent of its non-descendent, given its parents.

Bayesian belief networks are very influential tool for modelling causes and effects in a wide variety of domains. The Bayesian belief network (BBN) represents a set of random variables and their conditional independencies using a directed acyclic graph (DAG), in which nodes represent random variables and missing edges encode conditional independencies between the variables.

4.3 COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES

Data mining techniques discussed here have their intrinsic limitations and assumptions that make one technique better than the others. Data mining techniques can be compared on the basis of following performance criteria [Zhang04]:

a) Classification Accuracy: Indicates how accurately a technique classify a fraud or non-fraud organization
b) Ease of problem encoding: mainly concerns that how complex is a technique in encoding a problem
c) Flexibility: deals with ability to handle various data types and a wide range of problems
d) Computation complexity: calculates the cost involved in generating results
e) Interpretability: concerns about the ability to explain data mining results clearly
f) Optimization capability: pertains to generating optimal results
g) Scalability: refers to the amount of extra effort required by a data mining technique to obtain results from a large scale data set.
h) Accessibility: refers to the availability of software.

Data mining techniques discussed in this chapter can be compared on the basis of above mentioned performance criteria on a five point scale ranging from low to very high (Figure 4.2).
4.4 EFFECTIVENESS OF DATA MINING METHODS

Each data mining technique has some advantages and limitations. Neural networks are competent enough for management of inconsistent or noisy data and make no assumptions regarding attributes' independence. Neural network appears the best technique in terms of scalability. Genetic algorithm is found to be one of the best techniques for handling missing values in training data. Genetic algorithm suffers from the problem of high cost involved in generating the results. This limitation of genetic algorithm can be overcome by using Decision Tree, since moderate cost is involved. Decision trees provide a meaningful way of representing acquired knowledge.

Decision tree appeared as a complex technique in encoding a problem. This drawback of decision tree can be best handled by using a genetic algorithm or a Bayesian belief network. Bayesian belief network found to be the best technique in terms of classification accuracy.
4.5 SUMMARY

Data Mining is an iterative process and is primarily useful in a scenario in which there are no predetermined notions about an interesting outcome. The selection of data mining method is one of the key components of this process. The comparative analysis given in this chapter will help auditors / investigative units in selecting a proper data mining algorithm for differentiating organisations into fraud or non-fraud.

This chapter presents an introduction to four data mining techniques commonly used for detection of financial statement fraud. A brief introduction of extensively used data mining technique is given as Section 4.2. Section 4.3 performs a comparative analysis of four data mining techniques on the basis of eight performance criteria. These performance criteria includes: classification accuracy, ease of problem encoding, flexibility, computation complexity, interpretability, optimization capability, scalability and accessibility. On the basis of comparative analysis Section 4.4 presents a brief overview of effectiveness of data mining techniques. This analysis will help the user in selecting a better data mining technique for detecting fraudulent financial reporting.