International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-2 Issue-4 June-2012

Document Fraud Detection with the help of Data Mining and Secure Substitution Method with Frequency Analysis

Namrata Shukla[1], Shweta Pandey[2]
M-Tech, Computer Science[1], Professor, Computer Science [2]
Shri Vaishnav Institute of Technology & Science, Indore [1,2]

Abstract

Prevention of fraud and abuse has become a major concern of many organizations. The industry recognizes the problem and is just now starting to act. Although prevention is the best way to reduce frauds, fraudsters are adaptive and will usually find ways to circumvent such measures. Detecting fraud is essential once prevention mechanism has failed. Several data mining algorithms have been developed that allow one to extract relevant knowledge from a large amount of data like fraudulent financial statements to detect. In this paper we present an efficient approach for fraud detection. In our approach we first maintain a log file for data which contain the content separated by space, position and also the frequency. Then we encrypt the data by substitution method and send to the receiver end. We also send the log file to the receiver end before proceed to the encryption which is also in the form of secret message. So the receiver can match the data according to the content, position and frequency, if there is any mismatch occurs, we can detect the fraud and does not accept the file.

Keywords

Fraud, fraudulent financial statements, Substitution Method, Frequency

1. Introduction

Fraud is an intentional act meant to induce another person to part with something of value, or to surrender a legal right. It is a deliberate misrepresentation or concealment of information in order to deceive or mislead. Fraud can range from minor employee theft and unproductive behavior to misappropriation of assets and fraudulent financial reporting. In different situational contexts, fraud can take somewhat different forms for example, bribery, embezzlement, securities fraud, health care fraud, money-laundering scams, insurance fraud, software piracy, internet fraud, telemarketing fraud, mortgage foreclosure scams, and identity theft -- these all have their own special characteristics. There are at least as many types of fraud as there are types of people who commit it. But in each instance, fraud involves deception. Someone knowingly lies in order to obtain an unlawful benefit, or an unfair advantage.

Some examples of fraud include:

- Any dishonest or fraudulent act;
- Forgery or alteration of a check, bank draft, or financial document;
- Misappropriation of assets;
- Deliberate impropriety in the handling or reporting of money or financial transactions
- Wrongfully using influence in a business transaction to receive a benefit
- profiteering as a result of insider information;
- Disclosing insider information to another person in or Abuse is behavior that is deficient or improper when compared with behavior that a prudent person would consider reasonable and necessary business practice given the facts and circumstances. Instances of abuse are not fraud or illegal acts, but they are harmful, and they need to be minimized.

Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models.

Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns. Data mining is not new, it has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and Cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling.

Data mining is about finding insights which are statistically reliable, unknown previously, and actionable from data [1]. This data must be available, relevant, adequate, and clean. Also, the data mining

problem must be well-defined, cannot be solved by query and reporting tools, and guided by a data mining process model [2].
The term fraud here refers to the abuse of a profit organisation's system without necessarily leading to direct legal consequences. In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications. It is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms.

We provide here an overview of executing data mining services with fraud detection along with the security applications. The rest of this paper is arranged as follows: Section 2 introduces Data Mining in Fraud Detection; Section 3 describes about Fraud Types; Section 4 shows the Literature Review; Section 5 describes the algorithm and proposed work. Section 6 describes the Conclusion and outlook.

2. Data Mining in Fraud Detection

For fraud detection, the client currently had a distinct data repository where model scoring was performed. Based on the model score, reports would be queried against the data warehouse to produce the claims that were suspect. This process was inefficient for a number of reasons:

- Fraud detection analysis had to be conducted by a specialist who passed the scores onto the fraud detection team. The team performed the investigation and evaluated the merits of the claims. This was a disjointed process whereby the fraud detection data mining scores were not being improved based on investigation results.
- Fraud detection was not receptive to sudden changes in claim patterns. For natural disaster events, such as hurricanes, there would be a spike in similar claims. The data mining score would not be able to adapt to such scenarios.
- Data mining was confined to actuarial specialists and not day-to-day managers.

Initially DWreview examined the text mining solution provided by the major data mining vendors. While these proved to be rich in features, they were cumbersome to use and lacked the finesse for ongoing daily usage. It became soon apparent that it would not be cost effective to implement such solutions for the client's needs.

The customized solution was developed in three modules. A scripting engine was designed and developed for the data extraction layer. The data extraction layer was designed to pull reports, either manually or as a scheduled task, from the data warehouse repositories. The scripting language used by the data extraction module is Perl, which makes the extraction module highly accessible for making changes by end-users.

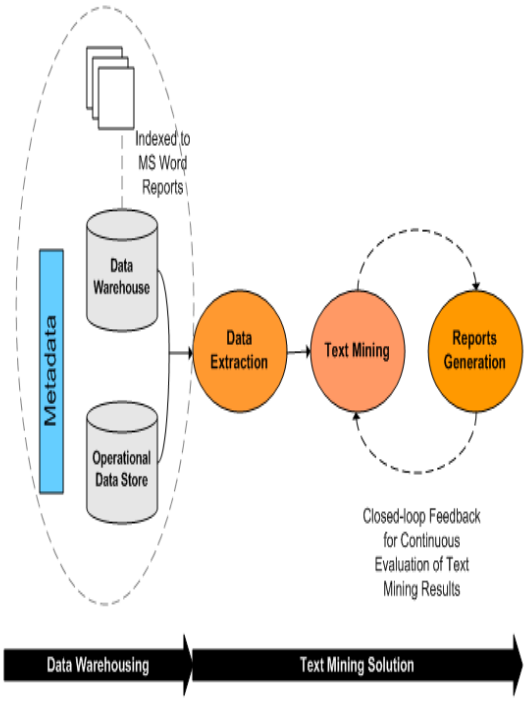

Figure 1: Text Mining Modules

The text mining module contains the data mining scores based on historically analysis of likelihood of fraud. The algorithms were custom developed based on text entered in the claims examiner's reports and details based on the claim. This model was developed specifically for the client by DWreview. The data mining model can give the client a competitive advantage and the technical details are kept as a closely guarded corporate secret.

3. Fraud Types

Types of fraud are shown in Figure 2.

TYPES OF FRAUD			
Type of Fraud	Perpetrator	Victim	Explanation
Employee embezzlement or occupational fraud	Employees	Employers	Employees directly or indirectly steal from their employers
Management Fraud	Top management	Stockholders, lenders, or others who rely on financial statements	Top management provides misrepresentation, usually in financial information
Investment scams	Individuals	Investors	Individuals trick investors into putting money into fraudulent investments
Vendor fraud	Organizations or individuals that sell goods or services	Organizations that buy goods or services	Organizations overcharge for good or services or no shipment of goods, even though payment is made
Customer Fraud	Customers	Organizations that sell goods or services	Customers deceive sellers into giving customers something they should not have or charging then less than they should

Figure 2: Types of Fraud

Financial crime poses a real and substantial threat to the stability of any business. Taking the proper measures to prevent or react quickly to a malfeasance is critical. Fraud and theft involving everything from intellectual property to inventory, from cybercrime to corruption, can be critically expensive.

PricewaterhouseCoopers' Global Economic Crime Survey 2009 found that the three most common types of economic crimes experienced in the last 12 months were asset misappropriation, accounting fraud and bribery and corruption.

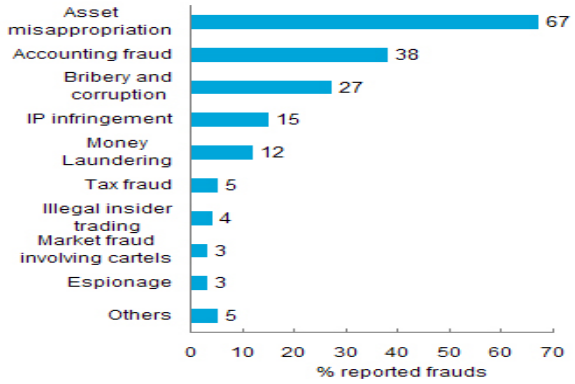

Figure 3: Types of economic crimes

The survey also shows that two-thirds of those respondents who have experienced economic crime in the last 12 months reported having suffered asset misappropriation. This type of fraud - the most prevalent since we began these surveys 10 years ago covers a variety of misdemeanors and while it is the hardest to prevent, it is arguably the easiest to detect. However, our 2009 survey shows that accounting fraud has become increasing prevalent. Of those respondents who reported economic crime in the past 12 months, 38% reported experiencing accounting fraud. This form of economic crime has significantly increased since 2007 and this appears to be linked to the economic cycle.

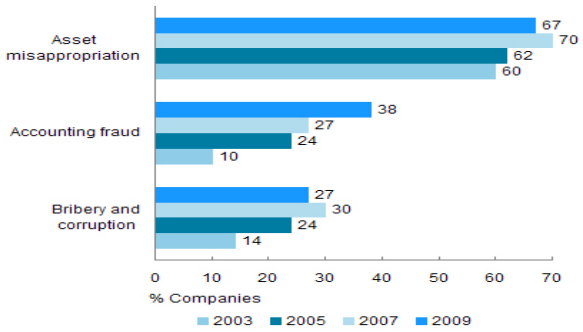

Figure 4: Trends in reported frauds

For organizations that encounter economic crime, fraud, or allegations of financial irregularity, our experienced and knowledgeable teams can manage and minimize the threat of corporate crimes and achieve improved outcomes using the following four front strategies:

- Reduce business disruptions, financial loss, and reputational damage;
- Identify the perpetrators and uncover actionable evidence;
- Trace and retrieve stolen/missing assets as fully as possible; and
- Recommend and/or implement effective remedial action to forestall future incidents.

We combine financial and accounting acumen with investigative skills and industry specific expertise to create a network of experts on an international level. Our global resources include CPAs, forensic accountants and Certified Fraud Examiners, former law enforcement officers, prosecutors, former financial regulators, lawyers, insurance specialists as well as

forensics technology and asset recovery specialists.

4. Literature Review

In 2005, Efstathios Kirkos et al. [3] explores the effectiveness of Data Mining (DM) classification techniques in detecting firms that issue fraudulent financial statements (FFS) and deals with the identification of factors associated to FFS. In accomplishing the task of management fraud detection, auditors could be facilitated in their work by using data mining techniques. This study investigates the usefulness of Decision Trees, Neural Networks and Bayesian Belief Networks in the identification of fraudulent financial statements. The input vector is composed of ratios derived from financial statements. The three models are compared in terms of their performances. The results identify the model with the best accuracy rate and highlight the importance of variables in fraudulent financial statement detection. They also indicate that the investigation of financial information can be of use in the identification of FFS and underline the importance of financial ratios.

In 2007, Dianmin Yue et al. [4] present a generic framework to guide our analysis. Critical issues for FSF detection are identified and discussed. Finally, they share directions for future research.

In 2009, G.Apparao et al. [5] analyzes that the prevention is the best way to reduce frauds, fraudsters are adaptive and will usually find ways to circumvent such measures. Detecting fraud is essential once prevention mechanism has failed. Several data mining algorithms have been developed that allow one to extract relevant knowledge from a large amount of data like fraudulent financial statements to detect FSF. It is an attempt to detect FSF; they present a generic framework to do our analysis.

In 2010, Shiguo wang et al. [6] categorizes, compares, and summarizes the data set, algorithm and performance measurement in almost all published technical and review articles in automated accounting fraud detection. Most researches regard fraud companies and non-fraud companies as data subjects, Eigen value covers auditor data, company governance data, financial statement data, industries, trading data and other categories. Most data in earlier research were auditor data; Later research establish model by using sharing data and public statement data. Company governance data have been widely used. It is generally believed that ratio data is more effective than accounting data; Seldom research on time Series Data Mining were conducted. The retrieved literature used mining algorithms including statistical test, regression analysis, neural networks, decision tree, Bayesian network, and stack variables etc.

In 2011, Tatsuya Minegishi et al. [7] focus on classification learning, which is an analytical method of stream mining. They are concerned with a decision tree learning called Very Fast Decision Tree learner (VFDT), which regards real data as a data stream. They analyze credit card transaction data as data stream and detect fraud use. In recent years, people with credit card are increasing. However, it also increases the damage of fraud use accordingly. Therefore, the detection of fraud use by data stream mining is demanded. However, for some data, such as credit card transaction data, contains extremely different rate of classes. Therefore, we propose and implement new statistical criteria to be used in a node-construction algorithm that implements VFDT. They also evaluate whether this method can be supported in imbalanced distribution data streams.

In 2011, Dr. Bhavani Thuraisingham [8] discuss the use of data mining for malicious code detection. Author introduces Data mining is the process of posing queries and extracting patterns, often previously unknown from large quantities of data using pattern matching or other reasoning techniques. Data mining has many applications in security including for national security as well as for cyber security. The threats to national security include attacking buildings, destroying critical infrastructures such as power grids and telecommunication systems. Data mining techniques are being investigated to find out who the suspicious people are and who is capable of carrying out terrorist activities. Cyber security is involved with protecting the computer and network systems against corruption due to Trojan horses, worms and viruses. Data mining is also being applied to provide solutions such as intrusion detection and auditing.

In 2012, Sherly K.K et al. [9] evaluates three classification methods to solve the fraud detection problems for data mining and shows how advanced techniques can be combined successfully to obtain high fraud coverage with maximum confidence and minimum false alarm rate.

In 2012, Clifton Phua et al. [10] observe that the credit application fraud is a specific case of identity crime. The existing nondata mining detection system of business rules and scorecards, and known fraud matching have limitations. To address these limitations and combat identity crime in real time, they propose a new multilayered detection system complemented with two additional layers: communal detection (CD) and spike detection (SD). CD finds real social relationships to reduce the suspicion score, and is tamper resistant to synthetic social relationships. It is the whitelist-oriented approach on a fixed set of attributes. SD finds spikes in duplicates to increase the suspicion score, and is probe-resistant for attributes. It is the attribute-oriented approach on a variable-size set of attributes. Together, CD and SD can detect more types of attacks, better account for changing legal behavior, and remove the redundant attributes. Experiments were carried out on CD and SD with several million real credit applications. Results on the data support the hypothesis that successful credit application fraud patterns are sudden and exhibit sharp spikes in duplicates. Although this research is specific to credit application fraud detection, the concept of resilience, together with adaptivity and quality data discussed in their paper, are general to the design, implementation, and evaluation of all detection systems.

5. Algorithm and Proposed work

Our proposed approach is shown in figure 3. In this approach we consider two types of Admin; one is of the sender side and other of the receiver side.

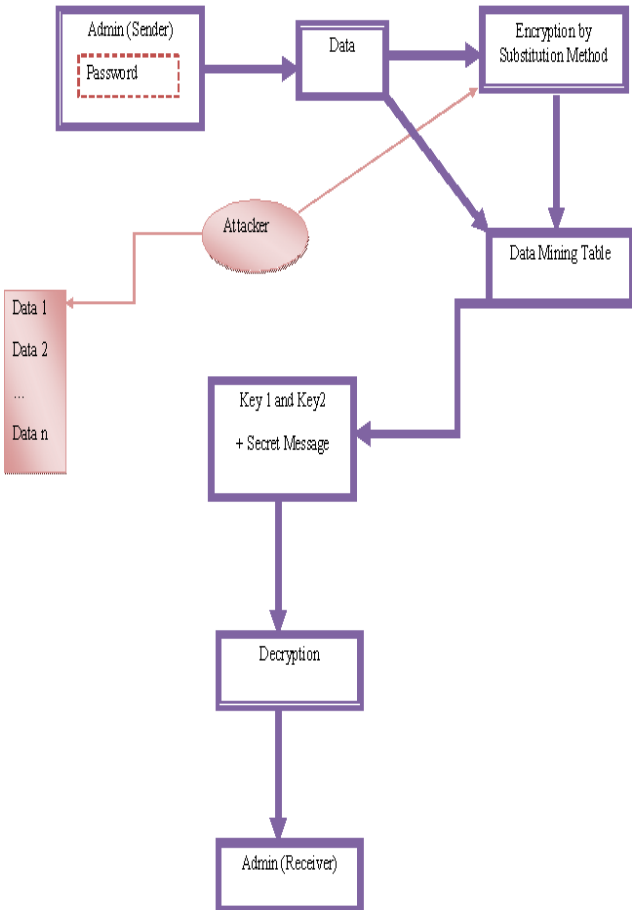

Figure 5: Proposed Approach

For explaning the encryption\decryption strategy we consider an example and coding for the alphabet which is shown in Figure 4.

A	B	C	D	E	F	G	H	I	J	K	L	M
0	1	2	3	4	5	6	7	8	9	10	11	12
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
13	14	15	16	17	18	19	20	21	22	23	24	25

Figure 6: A-Z coding Chart

In cryptography, a Caesar cipher, also known as a Caesar's cipher or the shift cipher or Caesar's code or Caesar shift, is one of the simplest and basic known encryption techniques.

It is a type of replace cipher in which each letter in the plaintext is replaced by a letter with a fixed position separated by a numerical value used as a

"key". Caesar Cipher is or was probably the very first encryption methodology.

Let me give an example:
Let's take a plain text: "DOTNETSPIDER IS A LEADING WEBSITE FOR DOT NET COMMUNITY" and we want to encrypt the plain text using "Caesar Cipher", using key/password as "9". Caesar Cipher only takes numeric password, and only between 0-25.

So our plain text's cipher/encrypted text will become: "MXCWNCBYRMNA RB J UNJMRWP FNKBRCN OXA MXC WNC LXVVDWRCH".

So how did we arrive at the cipher/encrypted text?
I will explain with a simple procedure.
Let us assign each alphabet in our English literature with a numeric value in a tabular format which is shown in figure 4.

Now take a word from the plain text, for example: "DOTNETSPIDER"

As I have said earlier that it is a type of replace cipher in which each letter in the plaintext is replaced by a letter with a fixed position separated by a numerical value used as a "key".
So we have to take each letter or character,In case of "DOTNETSPIDER": the value of 'D' is 3, so add 3 by 9 (since we have used '9' as our key/password)
So, 3 + 9 = 12
Now refer to the table above, now whose value is 12, its 'M'.

So here ‘D’ gets replaced by ‘M’
Let’s take another value from “DOTNETSPIDER” ? ‘T’: the value of ‘T’ is 19, so add 19 by 9 (since we have used ‘9’ as our key/password)
So, 19 + 9 = 28,
Here you can see the value is not on the list.
So whenever the value is greater than ‘25’, just subtract the value with ‘26’.
So, 19 + 9 = 28, and now the value is greater than 25, so, 28 – 26 = 2.
Now refer to the table above, now whose value is 2, its ‘C’. So here ‘T’ gets replaced by ‘C’

Similarly,
O (14) + 9 = 23(X)
T (19) + 9 = 28 – 26 = 2(C)
N (13) + 9 = 22(W)
E (4) + 9 = 13(N)
T (19) + 9 = 28 – 26 = 2(C)
S (18) + 9 = 27 – 26 = 1(B)
P (15) + 9 = 24(Y)
I (8) + 9 = 17(R)
D (3) + 9 = 12(M)
E (4)+ 9 = 13 (N)
R (17) + 9 = 26 – 26 = 0(A)
Hence, “DOTNETSPIDER” became, by using Caesar Cipher “MXCWNCBYRMNA”.

Same thing is applied with the numeric value.
- It would be better to use the following cipher:
- A → 26, …,X →03, Y →02, Z → 01 and space is 00
- We know that every TWO symbols represent a letter
- Thus
- 14260719001808000719220807
- is…
- MATH IS THE BEST

Figure 7 shows the stages of encryption and decryption.

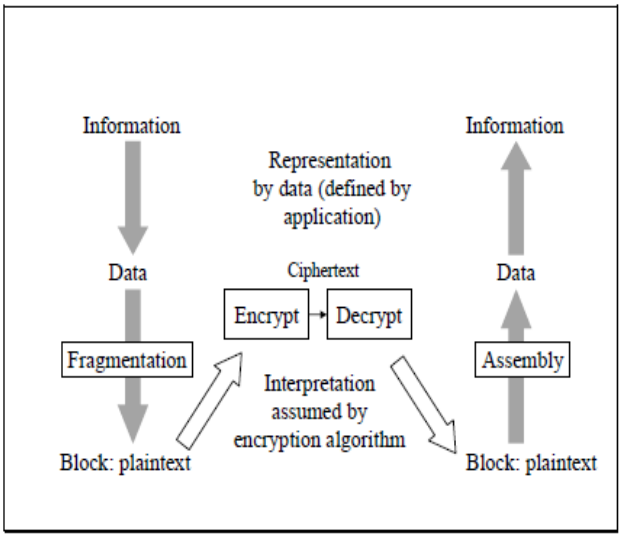

Figure 7: Encryption /Decryption

Algorithm: Substitution Method with Frequency Analysis
- In the Caesar cipher, the following algorithm is used
- If n is the number of a letter in the alphabet, this letter is replaced by another letter, whose number is (n+k) modulo 26 (shortly (n+k) mod 26)
- This is a remainder of division of (n+k) by 26

- For example, take k=5 and take letter X
- Its number n = 24
- (n+k) mod 26 = 24 + 5 mod 26 = 3
- So X is replaced with C
- Count the number of appearance of each letter and divide it by the total number of words in the ciphertext
- Compare the results with the frequency table
- Note: this method applies effectively to sufficiently large texts

If our file is attacked by intruder and it is successfully altered by the intruder, for this case we maintain a data mining log file for identify the fraud. Consider the example of Table 1, the message "Data Mining is data and mining" to be send to the Admin (Receiver). For this we maintain a log file which contain the content, position and frequency which will be separated by space. The same log file is sent to the receiver side [Table 2]. After successfully verification the receiver accepts the data otherwise receiver denies accepting the data or detecting the fraud. Same procedure will be applied to numeric data which is shown in Table 3 and Table 4 respectively.

Table1: String Message

Data Mining is data and mining.

Table 2: Log File

Content	Position	Frequency
Data	1,4	2
Mining	2,6	2
is	3	1
and	5	1

Table3: Numeric Message

15 2 15 6 19 15 6 15 2

Table4: Log File

Content	Position	Frequency
15	1,3,6,8	4
2	2,9	2
6	4,7	2
19	5	1

6. Conclusion and Outlook

Due to the extensive growth of E-Commerce fraud detection has become a necessity. The term fraud here refers to the abuse of a profit organizations system without necessarily leading to direct legal consequences. Fraud detection is a continuously evolving discipline and ever changing tactics to commit fraud.

In this paper we use the terminology of fraud by the hackers to mislead the information from the sender to the receiver. The main focus of this paper is to identify the attack from the hackers on any text file document which is encrypted and send to the receiver. For this identification we apply frequency based data mining technique to identify the fraud.

References

[1] Elkan, C. (2001). Magical Thinking in Data Mining: Lessons from CoIL Challenge 2000. Proc. of SIGKDD01, 426-431.

[2] Lavrac, N., Motoda, H., Fawcett, T., Holte, R., Langley, P. & Adriaans, P. (2004). Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving. Machine Learning 57(1-2): 13-34.

[3] Efstathios Kirkos, Charalambos Spathis, Yannis Manolopoulos," Detection of Fraudulent Financial Statements through the use of Data Mining Techniques", 2nd International Conference on Enterprise Systems and Accounting 2005July 11-12,Thessaloniki, Greece.

[4] Dianmin Yue, Xiaodan Wu, Yunfeng Wang, Yue Li, Chao-Hsien Chu," A Review of Data Mining-based Financial Fraud Detection Research", IEEE 2007.

[5] G.Apparao, Dr.Prof Arun Singh, G.S.Rao, B.Lalitha Bhavani, K.Eswar,D.Rajani," Financial Statement Fraud Detection by Data Mining", Int. J. of Advanced Networking and Applications ,Volume: 01 Issue: 03 Pages: 159-163 (2009).

[6] Shiguo wang," A Comprehensive Survey of Data Mining-based Accounting-Fraud Detection Research", 2010 International Conference on Intelligent Computation Technology and Automation.

[7] Tatsuya Minegishi, Ayahiko Niimi, "Detection of Fraud Use of Credit Card by Extended VFDT", IEEE 2011.

[8] Dr. Bhavani Thuraisingham," Data Mining for Malicious Code Detection and Security Applications", 2011 European Intelligence and Security Informatics Conference.

[9] Sherly K.K," A Comparative Assessment Of Supervised Data Mining Techniques for Fraud

Prevention",TIST.Int.J.Sci.Tech.Res.,Vol.1,2012, 1-6.

[10] Clifton Phua, Kate Smith-Miles,Vincent Cheng-Siong Lee, and Ross Gayler," Resilient Identity Crime Detection", IEEE Transactions On Knowledge and Data Engineering, VOL. 24, NO. 3, March 2012.