Computers in Human Behavior 28 (2012) 1002–1013

Contents lists available at SciVerse ScienceDirect

Computers in Human Behavior

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

Identifying the signs of fraudulent accounts using data mining techniques

Shing-Han Li [a,*], David C. Yen [b,1], Wen-Hui Lu [c,2], Chiang Wang [a,2]

[a] Department of Information Management, Tatung University, 40 ChungShan North Road, 3rd Section, Taipei 104, Taiwan
[b] Department of Decision Sciences and Management Information Systems, Miami University, Oxford, OH 45056, United States
[c] Department of Computer Science and Engineering, Tatung University, 40 ChungShan North Road, 3rd Section, Taipei 104, Taiwan

ARTICLE INFO

Article history:
Available online 4 February 2012

Keywords:
Fraud detection
Data mining
Dummy account
Fraudulent account
ATM phone scams

ABSTRACT

In today's technological society there are various new means to commit fraud due to the advancement of media and communication networks. One typical fraud is the ATM phone scams. The commonality of ATM phone scams is basically to attract victims to use financial institutions or ATMs to transfer their money into fraudulent accounts. Regardless of the types of fraud used, fraudsters can only collect victims' money through fraudulent accounts. Therefore, it is very important to identify the signs of such fraudulent accounts and to detect fraudulent accounts based on these signs, in order to reduce victims' losses. This study applied Bayesian Classification and Association Rule to identify the signs of fraudulent accounts and the patterns of fraudulent transactions. Detection rules were developed based on the identified signs and applied to the design of a fraudulent account detection system. Empirical verification supported that this fraudulent account detection system can successfully identify fraudulent accounts in early stages and is able to provide reference for financial institutions.

1. Introduction

As the improvement of modern civilization and technology has brought many new benefits to the society, it has also resulted in a transition of social values towards utilitarianism. This, in turn, leads to some people's chasing of increased wealth through fraudulent activities.

With the recent advancement of media and communication networks, many techniques to commit fraud have been developed, resulting in various new ways of fraud (Phua, Leem, Smith, & Gayler, 2010). ATM phone scams are one well-known type of fraud. Fraudsters use various scams to deceive and attract victims to transfer their money into fraudulent accounts[3] through the use of financial institutions or ATMs. There are many different types of scams, including lottery, tax refund, web shopping, kidnapping, fiduciary loan, and health insurance refund. Fraudsters use telecommunications technology to make scam calls from foreign regions, or use stolen identities (IDs) to apply for local phone numbers, in order to conceal their true identities and locations from the law enforcement. In addition, fraudsters need to use stolen IDs to apply for fraudulent accounts to receive money transfers from victims.

The government in Taiwan has realized the severity of these fraud crimes and took necessary procedures to stop frauds. According to the statistics announced by the National Police Agency (2011)[4], there were 28,820 cases of committed frauds, among which, telephone frauds took the largest portion (30.7%) with 8842 cases. Although the annual number of committed ATM phone scams was declining (National Police Agency, 2011), ATM phone scams remain harmful to victims and the entire society. As most ATM phone scams collect money from victims through fraudulent accounts, it is important for financial institutes to detect and identify fraudulent accounts to remove the tools from fraudsters.

There have been limited numbers of literatures related to ATM transfer fraud. The focuses of these prior researches include (1) statistical analysis related to fraud activities (Brentnall, Crowder, & Hand, 2008; Sudjianto et al., 2010); (2) process and technique analysis of fraud activities (Ku & Liao, 2007; Titus & Cover, 2001); (3) legal analysis of frauds (Hayhoe, 1995; White, 2008). This study intended to apply data mining techniques to assist the selected bank in identifying signs of fraudulent accounts from the vast transaction detail database and to detect fraudulent accounts based on these signs. The purpose of this study is to present new reference materials to detect fraudulent accounts in order to develop strategies for detecting and preventing future frauds.

Based on the motives above, this study has the following objectives:

* Corresponding author. Address: Department of Information Management, Tatung University, 40 ChungShan North Road, 3rd Section, Taipei 104, Taiwan. Tel.: +886 2 25925252x3610; fax: +886 2 25853966.
E-mail addresses: shli@ttu.edu.tw (S.-H. Li), yendc@muohio.edu (D.C. Yen), d9906007@ms.ttu.edu.tw (W.-H. Lu), u5510647@tknet.tku.edu.tw (C. Wang).
[1] Tel.: +1 513 229 4827; fax: +1 513 529 9689.
[2] Tel.: +886 2 25925252x3610; fax: +886 2 25853966.
[3] A fraudulent account is a dummy bank account used for the purpose of fraud.
[4] National Police Agency, Ministry of the Interior, Executive Yuan, ROC.

0747-5632/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chb.2012.01.002

(1) Identifying the signs of fraudulent accounts;
(2) developing a fraudulent account detection system;
(3) providing reference materials to detect fraudulent accounts in financial institutions, in order to decrease the probability of ATM phone scams and reduce the losses associated with such frauds.

2. Related research

2.1. ATM phone scams

Automatic Teller Machines, ATMs, provide services that include cash withdrawals, cash advances, transfers, tax payments, and balance inquiry. According to statistics of the Financial Supervisory Commission[5] (FSC, 2011), by March 2011, there were 25,745 ATMs and 156,260,000 ATM cards issued. The wide acceptance of ATM usage made it an easy target of fraud. Typically, fraudsters opened fraudulent accounts in financial institutes and made massive number of phone calls to home and mobile phone users to send fraudulent messages; victims deceived by these calls and messages then followed fraudsters' instructions to transfer their money into fraudulent accounts via ATMs (Federal Trade Commission, n.d). The steps of ATM phone scams are further explained in the following:

(1) Open fraudulent bank accounts:

The reliance on ID cards for applications of bank accounts, credit cards, and loans creates routes for various financial crimes. For ATM phone scams, the first, and the most common move is to apply for fraudulent accounts in financial institutions using stolen or fake IDs to hide their true identities (Administrative Enforcement Agency, 2007). These fraudulent accounts are usually used as temporary storage of victims' money before fraudsters' retrieval.

(2) Deceive victims using various scenarios:

An announcement from the Administrative Enforcement Agency (2007)[6] revealed the recent advancement of ATM phone scams by criminal organizations. Originally, fraudsters used fax, short message, or phones to inform victims about their wining of monetary or merchandize price. Once victims called back through the numbers they provided, fraudsters told victims that they need to pay tax for the price they won and convinced victims to transfer money to fraudulent accounts. With people's improved awareness of price-winning frauds, fraudsters changed their techniques to claim themselves as staff of government agencies or law enforcement, and inform victims about their personal accounts or IDs being used for frauds. Once victims bought the story, fraudster further convinced victims to transfer their money to fake custody accounts for guarding (Administrative Enforcement Agency, 2007).

(3) Utilize ATM to receive money:

Fraudsters typically instruct victims to transfer their money to fraudulent accounts using ATMs and withdraw the defrauded money through ATMs to escape ID confirmation from financial institutions. In addition, one particular technique used by fraudsters in Taiwan was to instruct victims to use English interface instead of Chinese interface, and with help of language barriers trick victims to transfer their money to fraudulent accounts (Administrative Enforcement Agency, 2007).

[5] Financial Supervisory Commission, Executive Yuan, ROC.
[6] Administrative Enforcement Agency, Ministry of Justice, Executive Yuan, ROC.

2.2. Using data mining in fraud detection

Data mining can be defined as a systematic process of exploring useful, potential information and features within large size of data (Rothman & Murphy, 1995). Academic research studies regarding data mining were diverse and covered a broad range of applications. Examples in business applications included customer behavior analysis (Au & Chan, 2003; Song, Kim, & Kim, 2001), marketing (Berry & Linoff, 1997; Shaw, Subramaniam, Tan, & Welge, 2001), customer management (Ngai, Xiu, & Chau, 2009), and business intelligence (Chen, Tsai, & Chang, 2008). As data mining provide methodology to collect, analyze, and identify useful data, the use of data mining in fraud detection helps in reducing the needs of manual screening (Phua et al., 2010).

Recently, various studies have been involved in financial fraud detection using data mining techniques, with the major focus on credit card related frauds. Leonard (1995) constructed a credit card fraud prediction model through the use of credit card authorization information from November 1991 to January 1992, including information of stores, transactions (place, time, amount, etc.), payments, credit card applications, and so on, as training data. Recursive algorithm was employed to divide training data sets for an expert system to generate rules. The generated rules were used in empirical test from February 1992 to March 1992, and results suggested that the model could detect 1000 potential fraudulent accounts per month with 50% fraud detection rate.

Ghosh and Reilly (1994) used data from Mellon Bank from January 1991 through June 1991 to develop a credit scoring system. With 30:1 ratio between legal and fraudulent accounts, a total of roughly 650,000 accounts were analyzed using neural network, and the credit scoring system was developed and installed in Melon Bank. The application of the credit scoring system successfully suppressed the occurrence of credit card fraud by 20–40% and reduced the burden of manual credit review (Ghosh & Reilly, 1994).

Kirkosa, Spathisb, and Manolopoulosc (2007) explored the effectiveness of data mining classification techniques in firms that issued fraudulent financial statements (FFS) and the identification of factors associated with these FFS. This study investigated the usefulness of Decision Trees, Neural Networks and Bayesian Belief Networks in the identification of fraudulent financial statements. Through the 10-fold cross validation, the authors concluded that Bayesian Belief Networks provided higher accuracy in fraud classification than the other two methods (Kirkosa et al., 2007).

Quah and Sriganesh (2008) proposed a real-time credit card fraud detection system. They presented a new and innovative approach to understanding spending patterns in order to decipher potential fraud cases. They made use of self-organization maps to decipher, filter, and analyze customer behavior for the detection of fraud (Quah & Sriganesh, 2008).

Several customer profiling studies have applied data mining on bank account analysis. The study of Au and Chan (2003) applied Fuzzy Association Rule Mining (FARM) to reveal hidden patterns in bank accounts to understand/identify different characteristics of the bank customers to provide better service and hence, to retain them. The FARM system employed both relational and transactional data of 320,000 bank customers and then, utilized various linguistic attributes to describe different patterns for rule establishment. Using the evaluation result from the identified bank's experts, 91.5% of the established rules were actually rated from the categories of useful to very useful (Au & Chan, 2003).

Ngai et al. (2009) conducted a review study in terms of the researches done in the subject area of the customer relationship management (CRM). The authors compared and contrasted

different data mining techniques used by prior research/studies and concluded that Neural Networks, Decision Tree, and Association Rules were the most frequently used data mining techniques in CRM field, while Decision Tree and Association Rules techniques were much easier to understand and be applied than Neural Networks in real applications (Ngai et al., 2009).

While Neural Networks and Decision Tree were two popular data mining techniques in CRM field, various studies have suggested that in terms of classification feature in data mining, Bayesian Classification showed better efficiency and effectiveness than Neural Network or Decision Tree method in some application areas such as pattern learning (Kirkosa et al., 2007; Mitchell, 1997; Phua et al., 2010).

According to these aforementioned literatures, no studies have been dedicated to focus on fraudulent account detection using data mining techniques. As fraudulent accounts play such a critical role in terms of ATM phone scams, it is important to develop a detection system to identify fraudulent accounts. In addition, data mining techniques are highly helpful in establishing patterns and rules for the development of detection system.

3. Research method

3.1. Selection of data mining techniques for ATM phone scam detection

The main objectives of this study were to explore and identify the actual transaction signs of fraudulent accounts, and to provide financial institutions and related policing-making sectors a reference in fraud prevention. As a fraudulent account is the major tools for fraudsters of ATM phone scams to collect and withdraw the defrauded money, it would be helpful if suitable data mining techniques can be used to identify the signs and reduce the number of these fraudulent accounts.

According to the study of Berry and Linoff (1997), data mining can be categorized into the following six modes including Classification, Estimation, Affinity Grouping, Sequence Pattern, Clustering, and Description. Two considerations were taken into account in the determination of data mining modes.

(1) As transactional details are categorical type of data by nature, classification is more suitable than estimation in general and in predicting the probabilities of variables.
(2) As fraudsters typically withdraw money from fraudulent accounts right after victims make their deposit or money transfer at ATMs, the association between a deposit and the subsequent withdrawal, such as the time difference and cash amount difference between these related transactions, is important.

Based on the above discussion, this proposed study intends to use Classification and Affinity Grouping (Association Rule) to perform data mining. As to the choice of Classification method, Bayesian Classification as per earlier discussion has provided a better efficiency in terms of pattern learning (Phua et al., 2010). For this reason, this proposed study selected Bayesian Classification and Association Rule to explore/investigate the information embedded within all transactional details. The two data mining techniques are briefly introduced as following:

(1) Bayesian Classification

The principle of Bayesian Classification originates from Bayesian Theorem, which calculates $P(H|X)$, the posterior probability that a condition H holds under in the sample X, through the following equation(Han & Kamber, 2006):

$$P(H|X) = [P(X|H)^{*}P(H)]/P(X),$$

where $P(X|H)$ is the posterior probability that sample X appears under the condition H. Further, $P(H)$ is the prior probability of H that is independent of sample X and $P(X)$ is the prior probability of X that is independent of H.

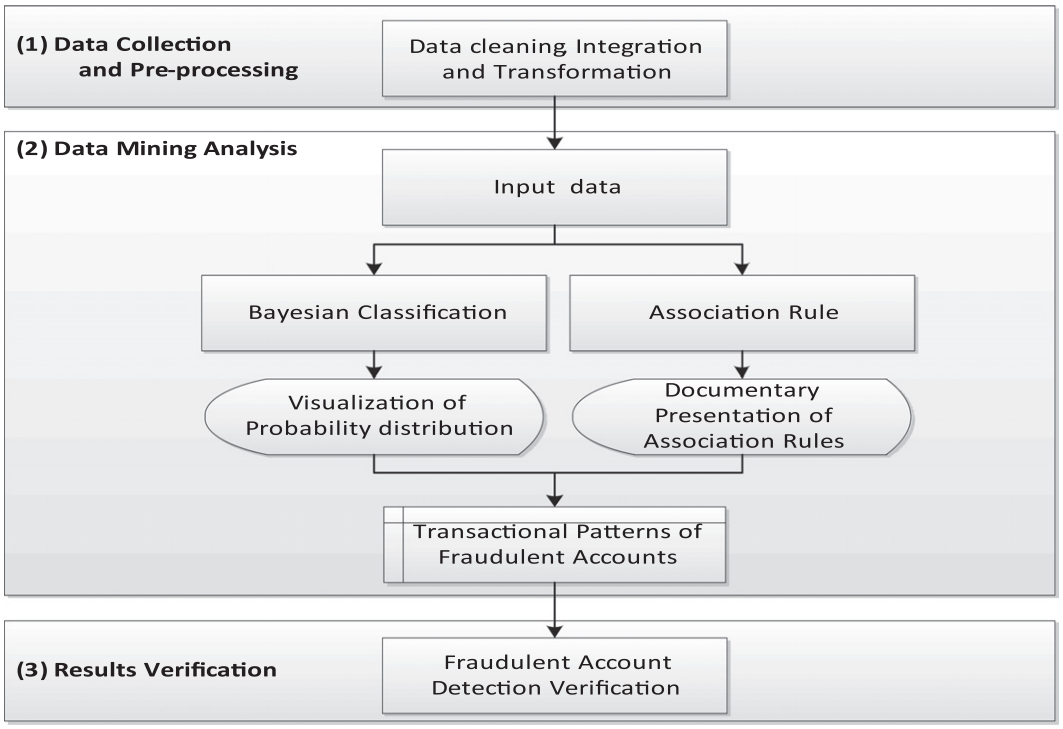

Fig. 1. Research process.

In this study, Bayesian Classification is used to analyze the customer account's details, the deposit transaction's details, and the withdrawal transaction's details using the training data set. Sample X refers to fraudulent or normal accounts, condition H refers to various attributes within the account details. The comparisons of the probabilities of attributes between fraudulent and normal accounts lead to the determination of the 'signs' of fraudulent accounts.

(2) Association Rule

Data mining using Association Rule to investigate the associated relationships in a given data set. The study of Agrawal, Imielinski, and Swami (1993) proposed the first Association Rule mining task to explore relationships among items (variables) in a supermarket basket transaction (data set) and two probabilities were considered as the measures to evaluate the strength of association between two independent variables:

$$\text{Support} = \sigma(X \cup Y)/N,$$

$$\text{Confidence} = \sigma(X \cup Y)/\sigma(X),$$

where N is the total number of data in the data set, $\sigma(X)$ is the number of data containing variable X, and $\sigma(X \cup Y)$ is the number of data containing both X and Y.

Support is the measure of the interestingness of an Association Rule because a rule with low support may be generated by chance.

Confidence, the conditional probability under the existence of X, measures the reliability of the Association Rule $X \rightarrow Y$. The higher confidence means that Y is more likely to appear in the data transaction containing X (Agrawal et al., 1993).

In this study, the probabilities were calculated for the time difference (X) and money amount difference (Y) between the consecutive deposit and withdrawal transactions occurred within fraudulent accounts.

3.2. Research process structure

As shown in Fig. 1, the research process structure of this study followed data mining procedure proposed by Feelders, Daniels, and Holsheimer (2000), with some modifications and result verification added. Repetitive verification and continuous improvement of procedure details and identified signs are important for the development of future fraudulent account detection. The following is a description of the research procedures:

(1) Data collection and pre-processing:

Data pre-processing is one of the most important steps in data mining. It ensures the correctness and the quality of the data used for data mining. In this study, the unprocessed bank account and transaction details were not suitable for direct data mining analysis. A COBOL 85 program was used to collect, integrate and transform the needed data from the bank's mainframe. As one of the earliest high-level programming languages, COBOL (COmon Business Oriented Language) was designed to support simple but large-amount business data processing and is now still widely accepted and used in the mainframe computers of some business sectors such as banking, finance, and accounting (Feelders et al., 2000). The selected bank in this study also used COBOL 85 to write/code programs to handle the task of database accessing. In order to reduce the cost and complexity in data collection and pre-processing, this study chose to comply with COBOL 85 to access, process, and export account details stored in the database of the selected bank.

(2) Data mining analysis:

As per earlier discussion, this study chose Bayesian Classification and Association Rule to perform data mining analysis. Bayesian Classification typically uses a small amount of training sample data to establish essential parameters for the succeeding data mining before analyzing the real data. Data analyzed using Bayesian Classification may include such items as the details of customer account, the deposit transaction details, and the details of withdrawal transactions. Association Rule is used to explore/investigate the association existed among various data fields. The related data signs are further identified by assigning attribute values to the variables and then, comparing the probabilities of different combinations of attributes values. In other words, the deposit-withdrawal differences file was analyzed using Association Rule. Bayesian Classification and Association Rule were used in different process of data mining and analyses task. There was no specific sequence existed for using these two data mining techniques.

(3) Results verification:

The established patterns or identified account signs have to be verified using the real data. This study developed a fraudulent account detection system using the established account signs. The detection system was mainly used for screening fraudulent accounts of the selected bank for around one-week period. Obtained results were then, compared with the fraudulent accounts reported by the Financial Supervisory Commission (FSC) in Taiwan.

4. Experiments and analysis

4.1. Data collection and pre-processing

This study used actual data from a selected bank in Taiwan, and the specific data sets used include customer account master files, customer name/address files, and customer transaction detail data files. The data collection and pre-processing process were shown in Fig. 2 and explained further in the following items:

(1) The transaction data for data mining analysis and training consisted of a sample of the transactions collected from the end of 2009 to early 2010, during which period a total of 327 fraudulent accounts of the bank were officially identified and reported by FSC. The sample accounts consisted of these 327 fraudulent accounts in mixture with 3,975,323 normal accounts.
(2) Following the study of Ghosh and Reilly (1994) which suggested the number of normal accounts should be limited to be approximately 30 times of fraudulent accounts in order to allow fraudulent accounts to show discernable signs, this study randomly selected 10,000 normal accounts for 327 fraudulent accounts. In addition, as the fraudulent accounts commonly used in ATM phone scams are mostly personal accounts, the selected sample size of normal accounts were further reduced to include the bank's personal accounts only. Finally, a total of 10,216 accounts, including 9889 normal accounts and 327 fraudulent accounts, formed the sample to perform the data mining task.
(3) A COBOL 85 program was used to preprocess the collected data. The customer account master file and the customer name/address file were first combined to form the personal account file (pbmr). Then, transaction details of customer accounts were extracted from the personal account file (pbmr) and separated into a deposit transaction detail file (htrmr_credit) and a withdrawal transaction detail file

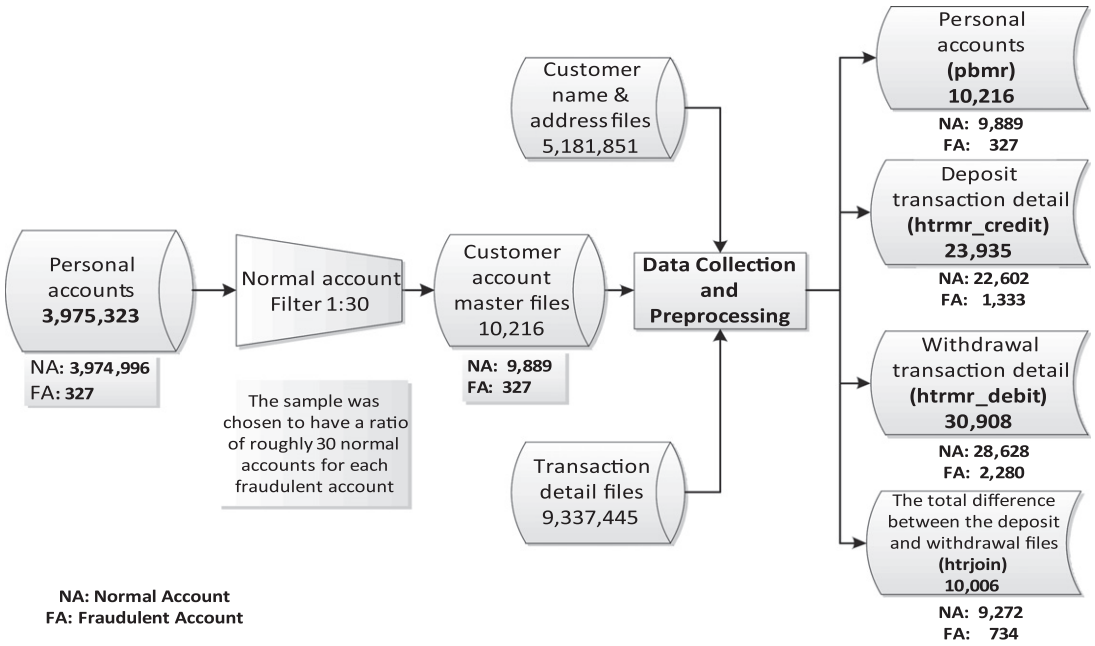

Fig. 2. Data collection and pre-processing process.

(htrmr_debit). As shown in Fig. 2, the total numbers of transactions details were 22,602 for normal accounts deposit, 1333 for fraudulent accounts deposit, 28,628 for normal account withdraw, and 2280 for fraudulent account withdraw.

(4) As bank transactions can be basically categorized into deposit and withdraw, the relationship between the deposit and withdraw is important. This study used the transaction detail files to calculate the differences in time and in amount of money between consecutive deposit and withdrawal transactions within the same account in order to generate the deposit-withdraw difference file (htrjoin). The total number of difference records was 9272 for normal accounts and 734 for fraudulent accounts, respectively.

(5) The detailed data field assignments of the personal account file, the deposit transaction detail file, the withdrawal transaction detail file, and the deposit-withdraw difference file were shown in Fig. 3.

4.2. Data mining analysis

This study applied Bayesian Classification and Association Rule to analyze and identify signs of fraudulent bank accounts. Data analyzed using Bayesian Classification included the customer account details, the deposit transaction details, and the withdrawal transaction details. The deposit-withdrawal differences file was analyzed using the Association Rule technique. The joint probability threshold of Bayesian Classification was set at 0.33 to detect attributes that displayed a high joint probability in the presence of fraudulent accounts (Flag 5 = 1).

Fig. 4 showed the results of Bayesian Classification of customer account file (pbmr). Age (Age) and contact phone number mark (Telflag) were identified to have a strong joint probability with fraudulent accounts (Flag 5 = 1).

Table 1 further compared the probability of attributes between fraudulent accounts and normal accounts. Related observations were described in the following:

- Observation 1 – Age:

The ages of persons that open fraudulent accounts were mostly (0.700) distributed between 21 and 40 (Age = 2). However, age 21–40 was also the group with the highest probability (0.447) for normal accounts. Therefore, age distribution cannot be used as a distinct sign for fraudulent accounts.

- Observation 2 – Contact Phone Number Information:

There existed a high probability (1.00) to have contact phone numbers included in the information of fraudulent accounts. However, contact phone numbers were included in both fraudulent and normal accounts (Telflag = 1). Therefore, whether an account includes contact phone number cannot be used as a distinct sign for fraudulent accounts.

Fig. 5 showed the results of Bayesian Classification of deposit transaction detail file (htrmr_credit). Inter-bank Transaction Mark (Cobkno), Transaction Abstract (Dscpt), Transaction Time (Txtime), and Transaction Type (Txtype) were identified to have a strong joint probability with fraudulent accounts (Flag 5 = 1).

Table 2 further compared the probability of the identified attributes between fraudulent accounts and normal accounts. Related observations were described in the following:

- Observation 3 – Inter-bank Transaction Mark:

The attribute value of Cobkno was 1 to mark the occurrence of inter-bank transaction, and 0 represented for no inter-bank transactions. Although a significant portion (0.348) of deposit transactions in fraudulent accounts were in the form of inter-bank transactions (Cobkno = 1), most deposit transactions were intra-bank transactions for both fraudulent (0.652) and normal accounts (0.899). Therefore, the inter-bank transaction mark in deposit transactions cannot be used as a distinct sign for fraudulent accounts.

- Observation 4 – Transaction Abstract:

According to the transaction abstract of each account, there existed a high probability (0.347) to deposit money into fraudulent

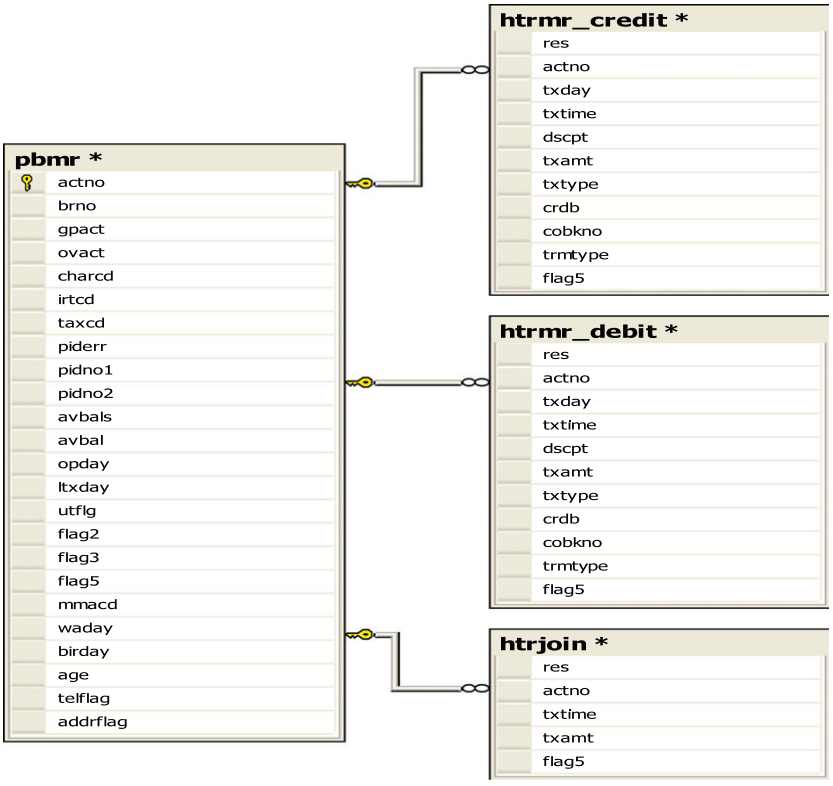

Fig. 3. Entity relationship diagram.

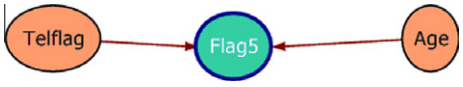

Fig. 4. Result of joint probability analysis of demographic attributes.

account through inter-bank transfer (Dscpt = MX). However, such a probability was not dominant enough to distinguish fraudulent and normal account. On the other hand, normal accounts had prearranged redeposit transactions (0.427) and salary-transfer deposits (0.073), while fraudulent accounts did not have prearranged redeposit transactions and salary-transfer deposits. Therefore, "no prearranged redeposit transactions" and "no salary transfer deposits" in deposit transactions can be used as distinct signs of potential fraudulent accounts.

- Sign 1: Fraudulent accounts have no prearranged redeposit transactions in the deposit transactions.
- Sign 2: Fraudulent accounts have no salary transfer deposits in the deposit transactions.

• Observation 5 – Transaction Time:

There existed a high probability (0.480) for deposit transactions in fraudulent accounts to occur during afternoon (Txtime = 2). On the other hand, a large portion (0.637) of deposit transactions in normal accounts took place in after hours (Txtime = 2). However, deposit transactions actually occurred during each of three time frames in both fraudulent accounts and normal accounts with a fairly even distribution. Therefore, transaction time in deposit transactions cannot be used as a distinct sign of fraudulent account.

• Observation 6 – Transaction Type:

There was a high probability (0.355) for deposit transactions in fraudulent accounts to occur in the form of ATM transactions. In contrast, the largest portion (0.687) of deposit transactions in normal accounts occurred in the form of transfer transactions. However, both fraudulent and normal accounts had certain portions of transfer transactions, cash transactions, and ATM transactions. In addition, while no batch transactions (Txtype = B) were found in fraudulent accounts, the probability of batch transactions to occur in normal accounts was also close to zero. Consequently, more features were needed to identify a distinct sign. From the above discussion, transaction type in deposit transactions cannot be used as a distinct sign of fraudulent accounts.

Fig. 6 showed the results of Bayesian Classification of withdrawal transaction detail file (htrmr_debit). Inter-bank Transaction Mark (Cobkno), Transaction Abstract (Dscpt), Transaction Amount (Txamt), Transaction Time (Txtime), and Transaction Type (Txtype) were all identified to have a strong joint probability with fraudulent accounts (Flag5 = 1).

Table 3 further compared the probability of the identified attributes between fraudulent accounts and normal accounts. Related observations were described in the following:

• Observation 7 – Inter-bank Transaction Mark:

There existed a strong probability (0.697) for withdrawal transactions in fraudulent accounts to take place through inter-bank transactions (Cobkno = 1). However, since both intra-bank transactions and inter-bank transactions occur in either fraudulent account or normal account, the inter-bank transaction mark in withdrawal transactions cannot be used as a distinct sign for fraudulent accounts.

Table 1
Comparison of the probabilities of attributes in customer account file (pbmr) between fraudulent and normal accounts.

Attribute	Account Numbers	Fraudulent account Probability (Flag5=1)	Normal Account Probability (Flag5=0)
Age=2 (21~40years)	4647	**0.700**	**0.447**
Age=3 (41~60years)	2402	0.153	0.238
Age=4 (over 60years)	2359	0.021	0.238
Age=1 (below 21years)	794	0.125	0.076
Telflag=1 (contact phone number included)	10216	**1.000**	**1.000**

Observation 1

Observation 2

- Observation 8 – Transaction Abstract:

According to the transaction abstract of withdrawal transactions, most (0.714) withdrawal transactions in fraudulent accounts were in the form of cash transaction (Dscpt = C). Although the similar dominance (0.450) of cash transactions in normal account withdrawals denied the use of cash withdrawals as a sign of fraudulent account, the zero probability of "collection and payment" transactions (Dscpt = ML) forms a distinct sign of fraudulent accounts.

– Sign 3: Fraudulent accounts have no "collection and payment" transactions in the withdrawal transactions.

- Observation 9 – Transaction Amount:

A significant large portion (0.953) of withdrawal transactions in fraudulent accounts had the transaction amount less than 130,000 dollars (Txamt < 13). However, an extremely similar distribution (0.901) existed in terms of the withdrawal transactions in the normal accounts. Therefore, transaction amount of withdraws cannot be used as a distinct sign of fraudulent accounts.

- Observation 10 – Transaction Time:

Transactions in fraudulent accounts typically occurred in the afternoon banking hours (0.366), or after hours (0.425). This probability ratio is similar to the one observed in the deposit transaction time of fraudulent accounts. This leads fact leads to a possible scenario whether or not a deposit and a withdrawal occurred in fraudulent account may have a close relationship in transaction time. However, since withdraws occurred in all three transaction time frames in both fraudulent and normal accounts, transaction time cannot be used as a distinct sign of fraudulent accounts.

- Observation 11 – Transaction Type:

In terms of transaction type, most (0.697) of the withdrawal transactions in fraudulent accounts were ATM transactions. Since withdrawal transactions in normal accounts were also concentrated (0.443) in the ATM transactions, ATM transactions (Txtype = A) cannot be used as the sign of fraudulent accounts. On the other hand, as shown in Table 3, fraudulent accounts did not have withdraws in the forms of batch transaction (Txtype = B) and on-line banking transaction (Txtype = 1), while normal accounts had a fairly large portion of batch withdrawal transactions and on-line banking withdrawal transactions. From the above discussion, "no batch transactions" and "no on-line banking transactions" in withdrawal transactions can be used as distinct signs of fraudulent accounts. Furthermore, this observation combined with the Observation 6 of deposit transactions in fraudulent accounts, it is clear that "no batch transactions" can be expanded to both deposit and withdrawal.

– Sign 4: Fraudulent accounts have no batch transactions in both deposit and withdrawal.
– Sign 5: Fraudulent accounts have no on-line banking transactions in withdrawal transactions.

As Bayesian Classification disclosed five distinct signs of fraudulent accounts, the other part of the data mining technique proposed in this study was focused on Association Rule analysis on the deposit-withdrawal difference file (htrjoin). Fig. 7 showed the result of Association Rule analysis of the differences between a deposit and withdrawal transactions. Two attributes, transaction time difference (Txtime) and transaction amount (Txamt) were used. The combination of time differences and amount differences between the associated deposit and withdraw were listed in sequence based on their joint probability in fraudulent accounts (Flag 5 = 1). Four values of Txtime (Txtime = 1 for within 15 min, Txtime = 2 for 15–30 min, Txtime = 3 for 30–45 min, and Txtime = 4 for 45–60 min time difference) and two values of Txamt (Txamt = 1 for NT$1000/US$34 and Txamt = 2 for NT$10,000/US$335) were shown in Fig. 7, with their corresponding joint probabilities. According to Fig. 7, most withdrawals in fraudulent accounts occurred within 60 min (Txtime = 1, 2, 3, and 4, with a total probability of 0.911) of the completion of previous deposits, and the amount difference between a deposit and a withdrawal within 60-min time difference was mostly within NT$10,000 (US$335; Txamt = 1 and 2). Therefore, a distinct sign was established for fraudulent accounts.

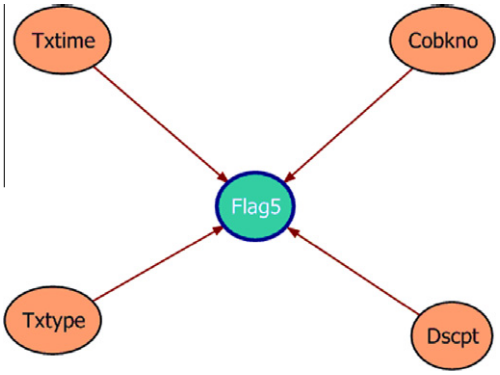

Fig. 5. Result of joint probability analysis of deposit transaction attributes.

Table 2
Comparison of the probabilities of attributes in deposit transaction detail file (htrmr_credit) between fraudulent and normal accounts.

Attribute	Account Number	Fraudulent account Probability(Flag5=1)	Normal Account Probability(Flag5=0)
Cobkno=0 (intra-bank transaction)	21177	0.652	0.899
Cobkno=1 (inter-bank transaction)	2758	**0.348**	**0.101**
Dscpt=XM (redepositing operation)	9655	**0.000**	0.427
Dscpt=C (cash transaction)	2930	0.213	0.117
Dscpt=MX (inter-bank transfer)	2097	**0.347**	0.072
Dscpt=MZ04 (salary transfer)	1656	**0.000**	0.073
Txtime=0 (outside banking hours)	14766	0.270	0.637
Txtime=2 (afternoon banking hours)	5160	**0.480**	0.200
Txtime=1 (morning banking hours)	4009	0.250	0.163
Txtype=M (transfer transaction)	16027	0.368	0.687
Txtype=C (cash transaction)	3073	0.253	0.121
Txtype=A (ATM TRANSACTION)	2365	**0.355**	**0.084**
Txtype=B (BATCH TRANSACTION)	1280	0.000	0.057

Observation 3
Sign 1
Observation 4
Sign 2
Observation 5
Observation 6

– Sign 6: The time and amount differences between associated deposits and withdrawals in fraudulent accounts are mostly occurred within 60 min and with a difference within NT 10,000 dollars.

From the above data mining analysis of fraudulent accounts, a total of 11 observations and 6 signs were collected. The identified signs, as listed in Table 4, can be used as reference for the subsequent detection of fraudulent accounts.

4.3. Experiment results verification

Early detection of fraudulent accounts shall result in reduced financial loss to victims and suppressed impact on the society. Accordingly, the verification method used in this study was to build a detection system based on the identified signs of fraudulent accounts and use daily transaction details to identify potential fraudulent accounts. The data used for verification was taken from the time period between APR 1 and APR 10, 2010, including APR 1, APR 2 and APR 5 thru APR 9 (seven business days). With APR 12 assigned as the checkpoint, the identified list of suspicious accounts was compared with the officially reported list of fraudulent accounts in the bank for verification.

Based on the identified six signs, a fraudulent account detection system was built using the combination of COBOL and WFL (Work Flow Language) programs. Fig. 8 summarized the process of verification of the system. The steps of the verification or the detection of fraudulent account were discussed in the following items:

(1) The collected account data was first screened to rule out salary transfer accounts and company accounts.
(2) Accounts having no transactions in one business day within the investigation period were further ruled out from the candidates list of fraudulent account for the same business day. The rest of the accounts were further screened using the identified six signs of fraudulent accounts.
(3) Accounts having a time difference within 60 min and an amount difference within NT 10,000 dollars between the associated deposits and withdraws (Sign 6) during the period of investigation were kept for further screening.
(4) Signs 1 through 5 were further screened in sequence with the inclusion of additional transactional details within a month prior to the investigation period. Accounts having no prearranged redeposit transactions (Sign 1), no salary transfer deposits (Sign 2), no "collection and payment" withdrawals (Sign 3), no batch transactions (Sign 4), and no on-line banking transactions were kept as the potential candidates of fraudulent accounts.
(5) The identified candidates of fraudulent accounts were gathered and compared/contrasted with the fraudulent accounts officially reported during the period of investigation.

Following the process of fraudulent account detection, 5245 potential fraudulent accounts were identified for further manual screening. Table 5 summarized the detection results in comparison with the fraudulent accounts officially reported by FSC.

An average of 749 suspicious accounts per day was obtained with 5245 potential fraudulent accounts identified during the period of 7 business days. Taking 749 as the average number of daily identified suspicious accounts, with the total number of approximately 160 branches, each branch in average needs to manually screen only 5 potential fraudulent accounts for each business

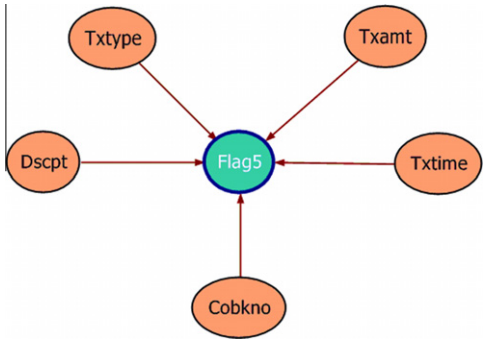

Fig. 6. Result of joint probability analysis of deposit transaction attributes.

Table 3
Comparison of the probabilities of attributes in withdrawal transaction detail file (htrmr_debit) between fraudulent and normal accounts.

attribute	Account Number	fraudulent accounts Probability(Flag5=1)	normal accounts Probability(Flag5=0)
Cobkno=0 (intra-bank transaction)	21480	0.303	0.726
Cobkno=1 (inter-bank transaction)	9428	**0.697**	0.274
Dscpt=C (cash transaction)	14508	**0.714**	0.450
Dscpt=ML (collection and payment)	266	**0.000**	0.093
Dscpt=MX (inter-bank transfer)	1400	0.041	0.046
Dscpt=M (normal transfer)	1314	0.011	0.045
Txamt <13 (10,000 transaction amount)	27957	0.953	0.901
Txamt=13~80 (10,000 transaction amount)	2165	0.036	0.073
Txamt=80~167 (10,000 transaction amount)	569	0.007	0.019
Txamt>=236 (10,000 transaction amount)	112	0.004	0.004
Txtime=0 (outside banking hours)	19259	**0.366**	0.644
Txtime=2 (afternoon banking)	6233	**0.425**	0.184
Txtime=1 (morning banking)	5416	0.209	0.173
Txtype=A (ATM transaction)	14267	**0.697**	0.443
Txtype=M (transfer transaction)	5730	0.018	0.199
Txtype=B (BATCH transaction)	5474	0.000	0.191
Txtype=1 (on-line banking transaction)	2412	0.000	0.084

Observation 7
Observation 8
Sign3
Observation 9
Observation 10
Observation 11
Sign 4
Sign5

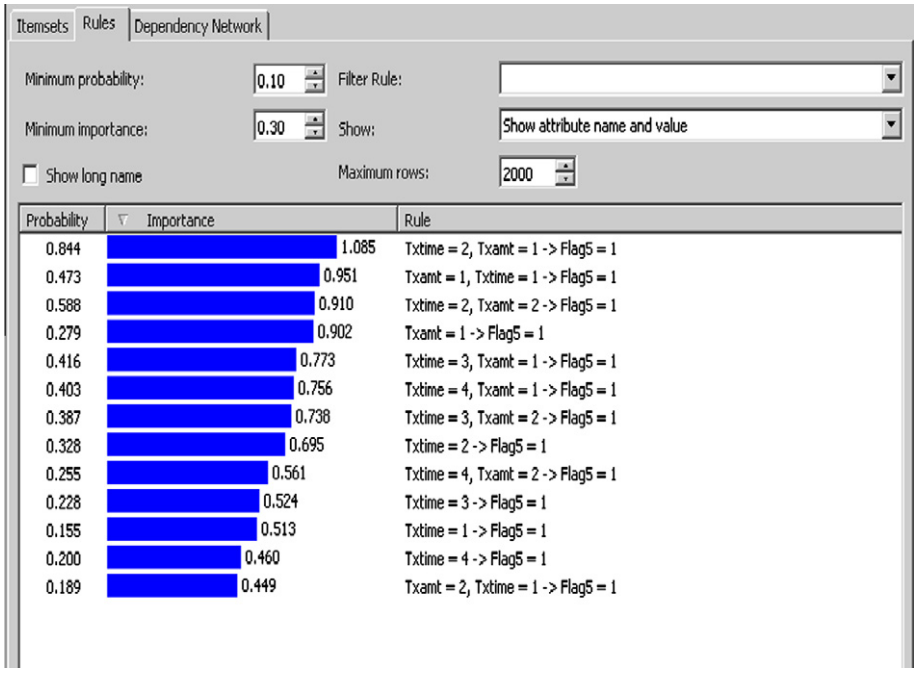

Fig. 7. The result of association rule analysis in deposit-withdraw difference file (htrjoin) for fraudulent accounts.

Table 4
Identified signs of fraudulent accounts.

Item #	Sign description
Sign 1	Fraudulent accounts have no prearranged redeposit transactions in deposit transactions
Sign 2	Fraudulent accounts have no salary transfer deposits in deposit transactions
Sign 3	Fraudulent accounts have no "collection and payment" transactions in withdrawal transactions
Sign 4	Fraudulent accounts have no batch transactions in both deposit and withdraw
Sign 5	Fraudulent accounts have no on-line banking transactions in withdrawal transactions
Sign 6	The time and amount differences between adjacent deposit and withdraw in fraudulent accounts are mostly within 60 min and NT 10,000 dollars

day. Such a requirement in human resources for the detection of fraudulent account is feasible.

In terms of the detection efficiency, as the number of fraudulent accounts is usually very small when compared with the total number of deposit accounts, it is inappropriate to assess the detection efficiency of fraudulent accounts using only the detection rate listed in Table 5. Considering the active accounts of 3,975,323 and fraudulent accounts of 327 during the period from late 2009 to early 2010 mentioned in Section 4, this study obtained the ratio of fraudulent accounts to total active accounts to be approximately 0.0082%. With the average detection rate of 0.400% in Table 5, if 327 fraudulent accounts are to be identified, the bank only needs to manually screen over 81,750 suspicious accounts and that is approximately 1/50 of the total active accounts. The efficiency of this fraudulent account detection system is indeed 50 times better than the efficiency of a purely manual screening.

As for the effectiveness of fraudulent account detection, during the period from APR 1 to APR 9, a total of 21 suspicious accounts were detected by the fraudulent detection system as fraudulent. After comparing the detailed information of the detected accounts, 8 repetitively accounts were identified and therefore the actual number of detected fraudulent accounts was 13. From APR 1 to APR 12, 20 actually fraudulent accounts were reported by FSC, while the fraudulent account detection system actually picked out 13 accounts and therefore the detection system's effectiveness of finding true fraudulent accounts can be calculated to be 65% (13/20). The actual effectiveness of fraudulent account detection should be higher

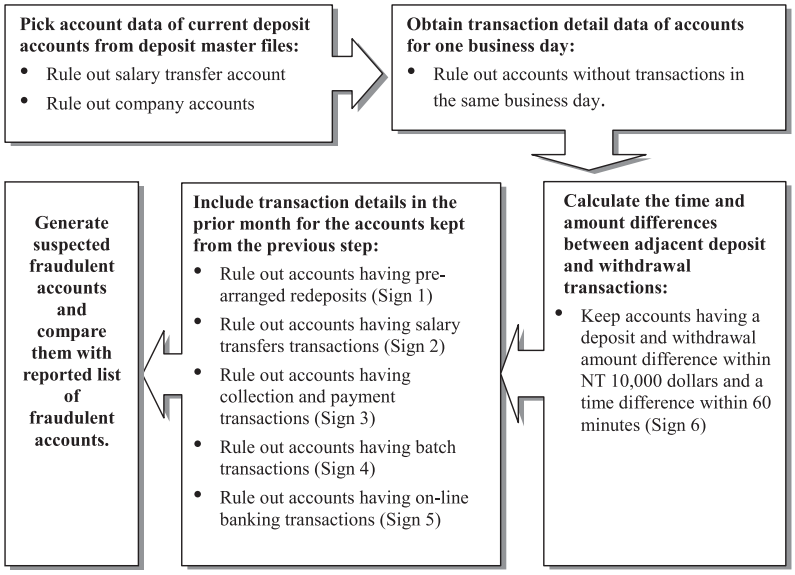

Fig. 8. The process of verification of the fraudulent account detection system.

Table 5
Results of Fraudulent Account Detection.

Date	Detected potential fraudulent accounts	Reported fraudulent accounts (by FSC)		Detection rate
APR 1	739	APR 2	1 fraudulent account	0.00541
		APR 6	1 fraudulent account	
		APR 7	2 fraudulent accounts	
APR 2	633	APR 2	1 fraudulent account	0.00473
		APR 5	2 fraudulent accounts	
APR 5	1213	APR 5	1 fraudulent account	0.00247
		APR 7	2 fraudulent accounts	
APR 6	648	APR 8	1 fraudulent account	0.00309
		APR 9	1 fraudulent account	
APR 7	680	APR 8	2 fraudulent accounts	0.00441
		APR 9	1 fraudulent account	
APR 8	573	APR 8	1 fraudulent account	0.00524
		APR 9	1 fraudulent account	
		APR 12	1 fraudulent account	
APR 9	759	APR 9	1 fraudulent account	0.00395
		APR 12	2 fraudulent accounts	
TOTAL	5245		21	0.00400

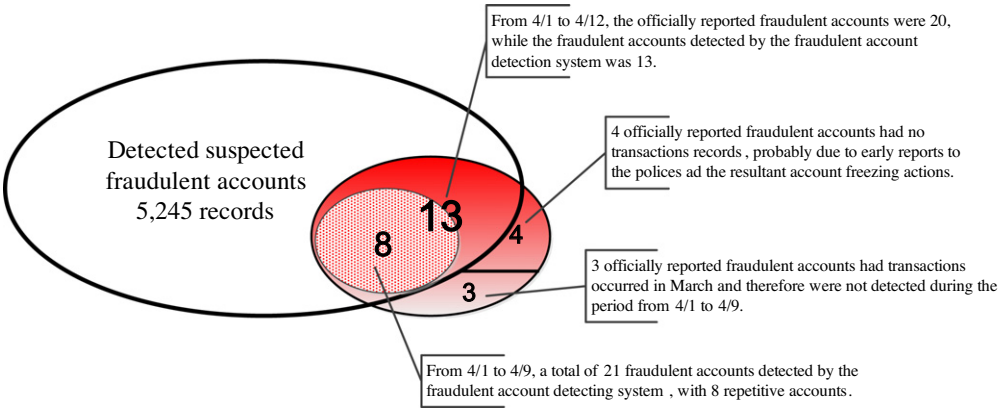

Fig. 9. Results of verification of the fraudulent account detection system.

than 65% since the deadline of the verification was limited to APR 12 and more detected suspicious accounts during the period might be reported as fraudulent accounts. A further analysis of the missing 7 reported fraudulent accounts from the detection system showed that the transactions of 3 reported fraudulent accounts occurred in March, beyond our verification period. The rest of 4 accounts did not have any transaction data, probably due to the early reporting action to the police and/or the resultant account freezing actions. Fig. 9 summarized the results of verification.

5. Conclusion

In conclusion, this study applied data mining techniques to analyze detailed daily account transaction data from a selected bank to establish detection signs for fraudulent account. A training data set consisting of 327 fraudulent accounts and 9889 normal accounts was pre-processed using a COBOL 85 program and Bayesian Classification and Association Rule techniques were applied to analyze transactional details to identify signs of fraudulent accounts. Six distinct signs were identified for the detection of fraudulent accounts. A fraudulent account detection system was developed based on these signs and the functionality of the fraudulent account detection system was further examined using real daily transaction data. Result of examination suggested that the fraudulent account detection method proposed by this study is more efficient and effective in detecting fraudulent accounts than purely manual screening.

Due to the characteristics of uncertainty, ambiguity, and complexity; the analysis of financial data may require a systematic methodology and a computerized technical support. Data mining was considered to be a highly suitable method for analyzing financial data. This proposed study applied data mining technique in the area of fraudulent account detection. The methodology and the results can be utilized to provide as a reference to future studies regarding the fraud detection. In addition, the fraudulent account detection system developed by this study was proven to be feasible through verification and hence can be adopted by financial institutions to reduce the need for a manual screening of fraudulent accounts.Furthermore, the major contributions of this study include the following items:

(1) Developing a fraudulent account detection system for COBOL-based banking systems.

As a large number of banking systems are usually COBOL-based, a fraudulent account detection system built upon COBOL programs can provide greater contribution to improve/enhance the resulting financial safety.

(2) Developing a fraudulent account detection system to assist the process of manual screening.

Instead of developing a fully automated fraudulent account detection system, this study provided a fraudulent account detection system to assist bank staffs to improve the efficiency and effectiveness to manual screen the fraudulent accounts while minimizing the impact on the increasing the structure and size of the work force and hence, adding the complexity of the bank. The development of such an assistive system also costs fewer resources and requires less time than that of a fully automated system.

(3) Providing preliminary standards to screen fraudulent accounts:

The 'signs' of fraudulent accounts obtained in this study can be used as the screening standards for screening in the development of a new fraudulent account detection system; in addition, the "observations" revealed by this study can be further utilized to develop more accurate screening standards in the development of a fully automated fraudulent account detection system. The implications of this study may include the following aspects.

(1) For academia:

While prior studies regarding data mining for fraud detection have been highly focused on credit card frauds, this study introduced a relative new research direction in applying data mining techniques in terms of fraudulent account detection. The verification of the effectiveness of the proposed fraudulent account detection system in fact, suggested that using data mining for fraudulent account detection is feasible and consequently, the proposed methodology and the obtained results of this study can be a starting point of future studies.

(2) For practitioners:

As fraudulent account detection is a new application area for data mining techniques, the development of commercial products fraudulent account detection system should be a good subject area for practitioners to explore/study. Either assistive or fully automated detection systems should find their application and implementation in various financial institutes.

(3) For users:

With fraudulent account detection system using data mining techniques proven to be feasible, financial institutes can make a plan to introduce such systems into their operations to better improve the productivity of their work force to concentrate better on

making the judgment of fraudulent accounts rather than focus on the data processing part. By doing so, the handling of fraudulent account detection can be proved to be more efficient and effective.The limitations of this study, however, may include the following items.

(1) Due to the concern of confidentiality, the detailed account information of the sample accounts was not presented in this study. The demonstration of detailed data processing procedures was also rather limited based on the same concern.
(2) Due to the difficulties associated with processing the enormous volume of bank transaction details, the size of the training data used in this study was limited to approximately 10,000 accounts having transactions during the period from late 2009 to early 2010;
(3) Results of this study were obtained based on the account transaction details of one bank and therefore the findings may not be able to be generalized to all financial institutions.

For future researches, data mining analysis based on transaction details of different banks or multiple banks should be helpful in generalizing the research results. Longitudinal analyses may also be helpful to understand the variation of transaction properties in different months, seasons, and years. Data mining with more sophisticated data transformation and classification is also a good research direction to pursue to compare/contrast the effect of data transformation and classification onto the obtained data mining results as well as the generated signs. Finally, a more detailed process and larger data set for the purpose of verification of fraudulent detection system may be another direction suitable for further investigation.

Acknowledgments

This work is partially supported by the National Science Council, Executive Yuan, ROC (NSC 100-2221-E-036-036) and Tatung University (B100-N02-053). The authors are also very appreciative for the helpful comments and suggestions provide by the editor and reviewers.

References

Administrative Enforcement Agency (2007). *Be careful of new tricks from fraudulent organizations* <http://www.tpk.moj.gov.tw/ct.asp?xItem=112289&ctNode=21720&mp=124>.
Agrawal, R., Imielinski, T., & Swami, A. (1993). Database mining: A performance perspective. *The IEEE Transactions on Knowledge and Data, 5*, 914–925.
Au, W. H., & Chan, K. C. C. (2003). Mining fuzzy association rules in a bank-account database. *IEEE transactions on fuzzy systems, 11*(2), 238–248.
Berry, M. J. A., & Linoff, G. S. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*. Hoboken: John Wiley & Sons, Inc.
Brentnall, A. R., Crowder, M. J., & Hand, D. J. (2008). A statistical model for the temporal pattern of individual automated teller machine withdrawals. *Journal of the Royal Statistical Society Series C, 57*(1), 43–59.
Chen, Y., Tsai, F. S., & Chang, K. L. (2008). Machine learning techniques for business blog search and mining. *Expert Systems with Applications, 35*(3), 581–590.
Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information Management, 37*(5), 271–281.
Financial Supervisory Commission (2011). *Automatic service machines of financial institutions. Financial statistics abstract* (pp. 89–91).
Ghosh, S., Reilly, D. L., (1994). Credit card fraud detection with a neural-network. In *Proceedings of the 27th annual Hawaii international conference on systems* (pp. 621–630).
Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques, 2/e*. Waltham: Morgan Kaufmann Publishers.
Hayhoe, R. (1995). Fraud, the consumer, and the banks: the (un-) regulation of electronic funds transfer. *University of Toronto Faculty of Law Review, 53*(2), 346.
Kirkosa, E., Spathisb, C., & Manolopoulosc, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*(4), 995–1003.
Ku, H. C., & Liao, H. C. (2007). The empirical analysis of internet fraud in taiwan—focusing on type of payment. *The Journal of Information Technology Society, 2007*(1), 37–52.
Leonard, K. J. (1995). The development of a rule based expert system model for fraud alert in consumer credit. *European Journal of Operational Research, 80*(2), 350–356.
Mitchell, T. (1997). *Machine Learning*. New York: The McGraw-Hill Companies, Inc.
National Police Agency (2011). *Fraud statistics*. <http://www.npa.gov.tw/NPAGip/wSite/public/Attachment/f1296196748375.doc>.
Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592–2602.
Federal trade commission (n.d.). *Recognize Phone Fraud*. <http://www.ftc.gov/bcp/edu/microsites/phonefraud/recognize.shtml>.
Phua, C., Leem, V., Smith, K., & Gayler, R. (2010). *A comprehensive survey of data mining-based fraud detection research* <http://arxiv.org/ftp/arxiv/papers/1009/1009.6119.pdf>.
Quah, J. T. S., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications, 35*(4), 1721–1732.
Rothman, M., & Murphy, E. (1995). *Data mining: a practical approach for database marketing*. Dallas: IBM White Paper.
Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *A Decision Support System, 31*, 127–137.
Song, H. S., Kim, J. K., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications, 21*, 157–168.
Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D., & Cela-Díaz, F. (2010). Statistical methods for fighting financial crimes. *Technometrics, 52*(1), 5–19.
Titus, R. M., & Cover, A. R. (2001). Personal fraud: The victims and the scams. *Crime Prevention Studies, 12*, 133–151.
White, E. E. (2008). Massively multiplayer online fraud: Why the introduction of real world law in a virtual context is good for everyone. *Northwestern Journal of Technology and Intellectual Property, 6*(2), 228.