ISSN:0975-9646

Data mining techniques for Fraud Detection

Anita B. Desai[#1], Dr. Ravindra Deshmukh[*2]

[#]*Sinhgad Institute of Management & Computer Application[#]*
Nahre Pune India
[2] *Ahmednagar College, Ahmednagar Dist-Pune India*

Abstract-Due to the dramatic increase of fraud which results in loss of billions of dollars worldwide each year, several modern techniques in detecting fraud are continually evolved and applied to many business fields. Fraud detection involves monitoring the behaviour of populations of users in order to estimate, detect, or avoid undesirable behaviour. Undesirable behaviour is a broad term including misbehaviour; fraud intrusion, and account defaulting. This paper present the concept of data mining and current techniques used in credit card fraud detection, telecommunication fraud detection, and computer intrusion detection. The goal of this paper is to provide a comprehensive review of data mining and different techniques to detect frauds.

Keywords: Fraud detection, computer intrusion, data mining, knowledge discovery, neural network

1. INTRODUCTION

Today, telecommunication market all over the world is facing a severe loss of revenue due to fierce competition and loss of income due to fraud. To survive in the market, telecom operators usually offer a variety of data mining techniques for fraud detection. According to telecom market, the process of subscribers (either prepaid or post paid) fraud continues to happen for any telecom industry, it would lead to the great loss of revenue to the company. . In this situation, the only remedy to overcome such business hazards and to retain in the market, operators are forced to look for alternative ways of using data mining techniques and statistical tools to identify the cause in advance and to take immediate efforts in response. This is possible if the past history of the customers is analysed systematically. Fortunately, telecom industries generate and maintain a large volume of data. They include Billing information, Call detail Data and Network data. This voluminous amount data ensures the scope for the application of data mining techniques in telecommunication database.

As plenty of information is hidden in the data generated by the telecom industries, there is a lot of scope for the researchers to analyze the data in different perspectives and to help the operators to improve their business in various ways. The most common areas of research in telecom databases are broadly classified into 3 types, i) Telecom Fraud Detection ii) Telecom Churn Prediction iii) Network Fault Identification and Isolation. Moreover, not all the data items of the telecom database are used by all the techniques. Only the relevant data items which really contribute to the specific analysis must be considered for any study. This study focuses on fraud detection the use of data mining techniques in fraud detection in telecomm data.

2. DATA MINING: AN OVERVIEW

2.1 Definition

"Data mining" is defined as a sophisticated data search capability that uses statistical algorithms to discover patterns and correlations in data [3]. Data mining finds and extracts knowledge ("data nuggets") buried in corporate data warehouses, or information that visitors have dropped on a website, most of which can lead to improvements in the understanding and use of the data.

Data mining discovers patterns and relationships hidden in data [4], and is actually part of a larger process called "knowledge discovery" which describes the steps that must be taken to ensure meaningful results. Data mining helps business analysts to generate hypotheses, but it does not validate the hypotheses

2.2. The evolution of data mining

Data mining techniques are the result of a long research and product development process. The origin of data mining lies with the first storage of data on computers continues with improvements in data access, until today technology allows users to navigate through data in real time. In the evolution from business data to useful information, each step is built on the previous ones. Table 1 shows the evolutionary stages from the perspective of the user.

In the first stage, Data Collection, individual sites collected data used to make simple calculations such as summations or averages.

The second step, company-wide policies for data collection and reporting of management information were established.

Table 1
Evolutionary stages of data mining

Stage	Business question	Enabling technologies	Product providers	Characteristics
Data Collection (1960s)	"What was my average total revenue over the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Navigation (1990s)	"What were unit sales in New England last March? Drill down to Boston"	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, IRI, Arbor, Redbrick, Evolutionary Technologies	Retrospective, dynamic data delivery at multiple levels
Data Mining (2000)	"What's likely to happen in Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Source: Pilot Software [17].

Once individual figures were known, questions that probed the performance of aggregated sites could be asked. For example, regional sales for a specified period could be calculated., a business could obtain either a global view or drill down to a particular site for comparisons with its peers (Data Navigation). Finally, on-

line analytic tools provided real-time feedback and information exchange with collaborating business units (Data Mining).

The core components of data mining technology have been developing for decades in research areas such as statistics, artificial intelligence, and machine learning.

2. 4. Data mining techniques

Data mining tools take data and construct a representation of reality in the form of a model. The resulting model describes patterns and relationships present in the data. From a process orientation, data mining activities fall into three general categories (see Fig. 1):

Discovery—the process of looking in a database to find hidden patterns without a predetermined idea or hypothesis about what the patterns may be.

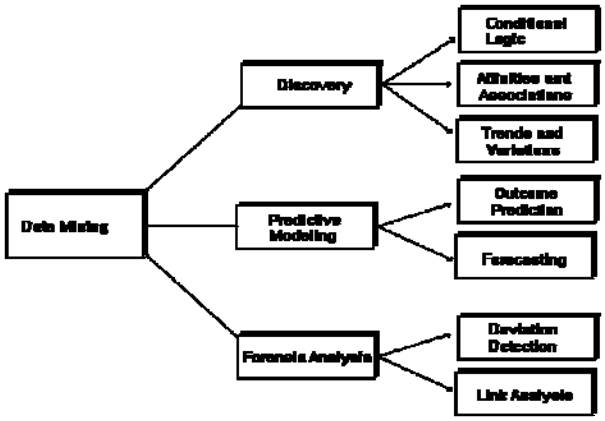

Fig. 1Breakdown of data mining from a process orientation. Source: Information Discovery, Inc. [18].

Predictive Modeling—the process of taking patterns discovered from the database and using them to predict the future.

Forensic Analysis—the process of applying the extracted patterns to find anomalous or unusual data elements.

Data mining is used to construct six types of models aimed at solving business problems: classification, regression, time series, clustering, association analysis, and sequence discovery [4]. The first two, classification and regression, are used to make predictions, while association and sequence discovery are used to describe behaviour. Clustering can be used for either forecasting or description.

Companies in various industries can gain a competitive edge by mining their expanding databases for valuable, detailed transaction information.

3 FRAUD : AN OVERVIEW

3.1 Definition

The Association of Certified Fraud Examiners (ACFE) defined fraud as "the use of one's occupation for personal enrichment through the deliberate misuse or application of the employing organization's resources or assets [1]." In the technological systems, fraudulent activities have occurred in many areas of daily life such as telecommunication net works, mobile communications, online banking, and E-commerce.

Fraud detection involves identifying fraud as quickly as possible once it has been perpetrated. Fraud detection methods are continuously developed to defend criminals in adapting to their strategies. The development of new fraud detection methods is made more difficult due to the severe limitation of the exchange of ideas in fraud detection At present, fraud detection has been implemented by a number of methods such as data mining, statistics, and artificial intelligence. Fraud is discovered from anomalies in data and patterns.

3.2 Type of Fraud

The types of frauds in this paper include credit card frauds, telecommunication frauds, and computer intrusion.

Credit Card Fraud.

Credit card fraud is divided into two types: offline fraud and online fraud. Offline fraud is committed by using a stolen physical card at storefront or call centre. In most cases, the institution issuing the card can lock it before it is used in a fraudulent manner. Online fraud is committed via web, phone shopping or cardholder- not-present. Only the card's details are needed, and a manual signature and card imprint are not required at the time of purchase.

Computer Intrusion

Intrusion is defined as the potential possibility of a deliberate unauthorized attempt to access information, manipulate information, or render a system unreliable or unusable. Intruders may be from an outsider (or hacker) and an insider who knows the layout of the system, where the valuable data is and what security precautions are in place.

Telecommunication Fraud

Fraud is costly to a network carrier both in terms of lost income and wasted capacity. The various types of telecommunication fraud can be classified into two categories: subscription fraud and superimposed fraud. Subscription fraud occurs from obtaining a subscription to a service, often with false identity details, with no intention of paying. Cases of bad debt are also included in this category. Superimposed fraud occurs from using a service without having the necessary authority detected by the appearance of unknown calls on a bill. This fraud includes several ways, for example, mobile phone cloning, ghosting (the technology that tricks the network in order to obtain free calls), insider fraud, tumbling (rolling fake serial numbers are used on cloned handsets so that successive calls are attributed to different legitimate phones), and etc.

4. DATA MINING TECHNIQUES IN FRAUD DETECTION

4.1.Credit Card Fraud Detection

Credit card fraud detection is quite confidential and is not much disclosed in public. Some available techniques are discussed as follows.

Outlier Detection. An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Unsupervised learning is a new explanation or representation of the observation data, which will then lead to improved future responses or decisions. Unsupervised methods do not need the prior knowledge of fraudulent and non-fraudulent transactions in historical database, but instead detect changes in behaviour or

unusual transactions. These methods model a baseline distribution that represents normal behaviour and then detect observations that show greatest departure from this norm. In supervised methods, models are trained to discriminate between fraudulent and non-fraudulent behaviour so that new observations can be assigned to classes. Supervised methods require accurate identification of fraudulent transactions in historical databases and can only be used to detect frauds of a type that have previously occurred. An advantage of using unsupervised methods over supervised methods is that previously undiscovered types of fraud may be detected.

Bolton and Hand proposed unsupervised credit card fraud detection, using behavioural outlier detection techniques[5].

Neural Networks. A neural network is a set of interconnected nodes designed to imitate the functioning of the human brain. Each node has a weighted connection to several other nodes in adjacent layers. Individual nodes take the input received from connected nodes and use the weights together with a simple function to compute output values, Neural networks can be constructed for supervised or unsupervised learning.

CARDWATCH [2] features neural networks trained with the past data of a particular customer. **It** makes the network process the current spending patterns to detect possible anomalies. Brause and Langsdorf proposed the rule- based association system combined with the neuro-adaptive approach [6]. Falcon developed by HNC uses feed-forward Artificial Neural Networks trained on a variant of a backpropagation training algorithm[11].

A neural MLP-based classifier is another example using neural networks II]. **It** acts only on the information of the operation itself and of its immediate previous history, but not on historic databases of past cardholder activities. A parallel Granular Neural Network (GNN) method uses fuzzy neural network and rule- based approach. The neural system is trained in parallel using training data sets, and then the trained parallel fuzzy neural network discovers fuzzy rules for future prediction. Cyber Source introduces a hybrid model, combining an expert system with a neural network to increase its statistic modeling and reduce the number of "false" rejections.

4.2 Computer Intrusion Detection

An intrusion detection system is needed to automate and perform system monitoring by keeping aggregate audit trail statistics. Intrusion detection approaches can be broadly classified into two categories based on model of intrusions: misuse and anomaly detection.

Misuese detection attempts to recognize the attacks of previously observed intrusions in the form of a pattern or a signature (for example, frequent changes of directory or attempts to read a password file) and directly monitor for the occurrence of these patterns . Misuse approaches include expert systems, model-based reasoning, state transition analysis, and keystroke dynamics monitoring . Misuse detection is simpler than anomaly detection. However, a primary drawback of misuse detection is that it is not possible to anticipate all the different attacks because it looks only known patterns of abuse.

Anomaly detection tries to establish a historical normal profile for each user, and then use sufficiently large deviation from the profile to indicate possible intrusions . Anomaly detection approaches include statistical approaches, predictive pattern generation, and neural net- works. The advantage of anomaly detection is that it is possible to detect novel attacks against systems. The techniques used in misuse detection and anomaly detection are described as follows:

Expert Systems. An expert system is defined as a computing system capable of representing and reasoning about some knowledge-rich domain with a view to solving problems and giving advice [16]. Expert system detectors encode knowledge about attacks as if-then rules. NIDES developed by SRI uses the expert system approach to implement intrusion detection system that performs real-time monitoring of user activity. NIDES consists of statistical analysis component for anomaly detection and rule- based analysis component for misuse detection.

Neural Networks. NNID (Neural Network Intrusion Detector) is an anomaly intrusion detection system implemented by a backpropagation neural network under UNIX environment[20]. ANN (Artificial Neural Networks) provides the ability to generalize from previously observed behaviour (normal or malicious) to recognize similar future unseen behaviour for both anomaly detection and misuse detection[10]. It is implemented by a backpropagation neural network.

Model-based Reasoning. Model-based detection is a misuse detection technique that detects attacks through observable activities that infer an attack signature. Garvey and Lunt combined models of misuse with evidential reasoning[9]. A pattern matching approach based on Colored Petri Nets to detect misuse intrusion is proposed by Kumar and Spafford [14]. **It** uses audit trails as input under UNIX environment.

Data mining approaches can be applied for intrusion detection. An important advantage of data mining approach is that it can develop a new class of models to detect new attacks before they have been seen by human experts. Classification model with association rules algorithm and frequent episodes is developed for anomaly intrusion detection [15]. This approach can automatically generate concise and accurate detection models from large amount of audit data.

State Transition Analysis is a misuse detection technique, which attacks are represented as a sequence of state transitions of the monitored system. STAT(State Transition Analysis Tool) is a rule-based expert system designed to seek out known

penetrations in the audit trails of multi-user computer systems [13). USTAT (UNIX State Transition Analysis Tool) is a UNIX-specific prototype of STAT . Other Techniques. ***A genetic algorithm*** [7] is applied to detect malicious intrusions and separate them from normal use. Dokas and Ertoz proposed building rare class predictive models for identifying known intrusions.

4.3. Telecommunication Fraud Detection

Most techniques use Call Detail Record data to create behaviour profiles for the customer, and detect deviations from these profiles. These approaches are discussed as follows.

Rule-based Approach.: A combination of absolute and differential usage is verified against certain rules in the rule based approach mapped to data in toll tickets[17]. Rule-based approach works best with user profiles containing explicit information, where fraud criteria can be referred as rules. Rule-discovery methodology combining two data levels, which are the customer data and behaviour data (usage characteristics in a short time frame)[19]. PDAT is a rule-based tool for intrusion detection developed by Siemens ZFE. Due to its flexibility and broad applicability

Neural Networks. Neural Networks can actually calculates user profiles in an independent manner, thus adapting more elegantly to the behaviour of the various users. A project of the European Commission, ASPeCT, investigated the feasibility of the implementations with a rule- based approach and neural networks approach, both supervised and unsupervised learning based on data in toll tickets[17].

Visualization Methods. Visualization techniques rely on human pattern recognition to detect anomalies and are provided with close-to-real-time data feeds. Visual data mining, combining human detection with machines for greater computational capacity, is developed by building a user interface to manipulate the graphical representation of quantities of calls between different subscribers in various geographical locations in order to detect international calling fraud [8].

Other Techniques. A call-based on-line fraud detection system based on a hierarchical regime-switching model is implemented by using subscriber data from real mobile communication network [12]. the gradient-based discriminative training is used to improve the performance. Location awareness of the mobile phone can be used to detect cellular clones within a local system and to detect roamer clones [18].

5. CONCLUSION

This paper begins with an overview of the concepts of data mining and fraud detection, followed by a discussion of evolution, characteristics, techniques. It also includes fraud detection in three areas, credit card fraud detection, computer intrusion detection, and telecommunication fraud detection . It presents the characteristics of fraud types, the need of fraud detection systems, several current fraud detection techniques. Most telecommunication fraud detection techniques explore data set of toll tickets and detect fraud from call patterns. These systems are effective against several kinds of frauds.

REFERENCES

[1] *Investigating Fraudulent Acts, UNIVERSITY OF HOUS-TON SYSTEM ADMINISTRATIVE MEMORANDUM.* http://www.uhsa.uh.edufsarn/AM/01C04.htm. 2000.

[2] E. Aleskerov, B. Freisleben, and B. Rao. Cardwatch: a neural network based database mining system for credit card fraud detection. In *Proceedings of Computational Intelligence for Financial Engineering,* pages 173-200, 1997. . .

[3] Newton's Telecom Dictionary, Harry Newton, CMP Books, http://www.cmpbooks.com

[4] Edelstein H. Data mining: exploiting the hidden trends in your data. DB2 Online Magazine. http://www.db2mag.com/9701edel.htm

[5] R. J. Bolton and D. 1. Hand. Unsupervised profiling meth- ods for fraud detection. In *conference of Credit Scoring and Credit Control VII. Edinburgh, UK. Sept* 5-7,2001.

[6] R. Brause, T. Langsdorf, and M. Hepp. Credit card fraud de- tection by adaptive neural data mining. In *Proceedings of the 11th IEEE International Conference on Tools with Artifi- cial Intelligence,* pages 103-106, 1999.

[7] A. Chittur, Model generation for an intrusion detection sys- tem using genetic algorithms. In *Ossining High school Hon- ors Thesis, 2001.*

[8] K. C. Cox, S. G. Eick, G. 1. Wills, and R.]. Brachman. Visual data mining: Recognizing telephone calling fraud. *J Data Mining and Knowledge Discovery,* 1(2):225-231, 1997.

[9] T. D. Garvey and T. F. Lunt. Model based intrusion detec- tion. In *Proceedings of the 14th National Computer Security Conference, October J99J.*

[10] A. K. Ghosh and A. Schwartzbard. A study in using neural networks for anomaly and misuse detection. In *Proceedings of the 8th USENlX Security Symposium, D.* c., 1999.

[11] K. Hassibi. Detecting payment card fraud with neural net- works. In *Business application of Neural Networks, P.J.G. Lisboa, A. Vellido, B.Edisbury Eds. Singopore: World Scien- tific, 2000.*

[12] J. Hollmn and V. Tresp. Call-based fraud detection in mo- bile communication networks using a hierarchical regime- switching model. In *Proceedings of the* 1998 *conference on Advances in neural information processing systems II,* pages 889-895. MIT Press, 1999.

[13] K. Ilgun, R. A. Kemmerer, and P. A. Porras. State transition analysis: A rule-based intrusion detection approach. *Soft- ware Engineering,* 21(3): 181-199, 1995.

[14] S. Kumar and E. H. Spafford. A Pattern Matching Model for Misuse Intrusion Detection. In *Proceedings of the 17th Na- tional Computer Security Conference,* pages 11-21, 1994

[15] w. Lee and S. Stolfo. Data mining approaches for intrusion detection. In *Proceedings of the 7th USENIX Security Sym- posium,* San Antonio, TX, 1998.

[16] T. F. Lunt, A. Tamaru, F. Gilham, R. Jagannathan, P. G. Neu- mann,H. S. Javitz, A. Valdes, and T. D. Garvey. AReal-Time Intrusion Detection Expert System (IDES) - Final Technical Report. Technical report, SRI Computer Science Laboratory, SRI International, Menlo Park, CA, Feb. 1992.

[17] Y. Moreau, B. Preneel, P. Burge, J. Shawe-Taylor, C. Stoer- mann, and C. Cooke. Novel techniques for fraud detection in mobile telecommunication networks. In *ACTS Mobile Sum- mit,* Grenada, Spain, 1997.

[18] S. Patel. Location identity and wireless fraud detection. In *ICPWC'97 Technical Program, Lucent technologies, Wire- less Secure Communications Lab, 1997.*

[19] S. Rosset, U. Murad., E. Neumann, Y. Idan, and G. Pinkas. Discovery of fraud rules for telecommunicationslchallenges and solutions. In *Proceedings of the fifth ACM SlGKDDinternational conference on Knowledge discovery and data mining,* pages 409-413. ACM Press, 1999.

[20] J. Ryan, M.-J. Lin, and R. Miikkulainen. Intrusion detec- tion with neural networks. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Pro- cessing Systems,* volume 10. The MIT Press, 1998.