Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Mining for Fraud Detection

[1]Manjunath K.V, [2]Patharaju S.D
[1]Samvardhana Coaching Centre, Bangalore, India
[2]SAMBA Bank, Riyadh, KSA

Abstract— Fraud detection is a topic applicable to many industries including banking and financial sectors, insurance, government agencies, telecommunication and law enforcement, and more. Fraud attempts have seen a drastic increase in recent years, making fraud detection is essential and more important than ever. Despite efforts on the part of the affected institutions, hundreds of millions of dollars are lost to fraud every year (leading to financial crisis and terrorism financing). Since relatively few cases show fraud in a large population, finding these can be tricky. In banking, (transactions involving fund transferred to/from the account is the primary source of data for AML). In insurance, 25% of claims contain some form of fraud, resulting in approximately 10% of insurance payout dollars. Fraud can range from exaggerated losses to deliberately causing an accident for the payout. With all the different methods of fraud, finding it becomes harder still. Data mining and statistics help to anticipate and quickly detect fraud and take immediate action to minimize costs. Through the use of sophisticated data mining tools, millions of transactions can be searched to spot patterns and detect fraudulent transactions.

Keywords— Data mining, Fraud detection, Supervised Learning, Unsupervised Learning, Clustering, classification.

I. INTRODUCTION

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. The advancement in technology and communication has created new opportunities for committing fraudulent acts. These acts impose serious threat to organizations on the financial, operational and psychological levels. In addition to the monetary losses, fraud can have a staggering effect on the organization's reputation, goodwill and customer relations. Therefore, organizations try to implement a variety of techniques to detect and prevent fraud. Among those techniques is data mining.

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. Data mining is about discovering new patterns which are unknown before, statistically reliable and process able from data. Data mining is a field (an area) which is concerned to understanding data patterns from huge datasets. We can say that the aim is to find out new patterns in data. A number of data mining techniques like Artificial Intelligence, classification, clustering, advanced neural networks, prediction and regression models used for different data mining approaches in various areas. Another area we are discussing here is fraud detection.

Fraud detection is the identification of symptoms of fraud where no previous disbelief exists. Firstly we have to learn that given data pattern is fraudulent or not. There are two kinds of learning data set supervised and unsupervised. Supervised learning of data set deals with fraud data that is previously known and unsupervised learning of data set deals with fraud data that is not previously considered as a fraud data but after sometimes they reflect the nature of fraud or crime. Then we treat those data patterns according to their behavior. Different terms are used for doing that task and they are described as techniques and methods for fraud or crime detection.

II. FRAUD

Fraud is an act of deception intended for personal gain or to cause a loss to another party. There are many words used to describe fraud: Scam, con, swindle, extortion, sham, double-cross, hoax, cheat, ploy, ruse, hoodwink, confidence trick. Fraud involves one or more persons who intentionally act secretly to deprive another of something of value, for their own benefit. Fraud is as old as humanity itself and can take an unlimited variety of different forms. However, in recent years, the development of new technologies has also provided further ways in which criminals may commit fraud. In addition to that, business reengineering, reorganization or downsizing may weaken or eliminate control, while new information systems may present additional opportunities to commit fraud.

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
- Identity theft

These 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.

III. FRAUD DETECTION

Fraud Detection involves finding the "needles in a haystack," which requires methodologies and techniques that are unique to this application area. Often, there are instances of fraud that have not been detected, which adds to the challenge of training the computer to recognize fraudulent cases. A complete fraud detection solution uncovers patterns of suspicious behaviour and provides actionable alerts to the organization.

In banking, fraud can involve using stolen credit cards, forging checks, misleading accounting practices, etc. In insurance, 25% of claims contain some form of fraud, resulting in approximately 10% of insurance payout dollars. Fraud can range from exaggerated losses to deliberately causing an accident for the payout. With all the different methods of fraud, finding it becomes harder still.

Data mining and statistics help to anticipate and quickly detect fraud and take immediate action to minimize costs. Through the use of sophisticated data mining tools, millions of transactions can be searched to spot patterns and detect fraudulent transactions. An important early step in fraud detection is to identify factors that can lead to fraud.

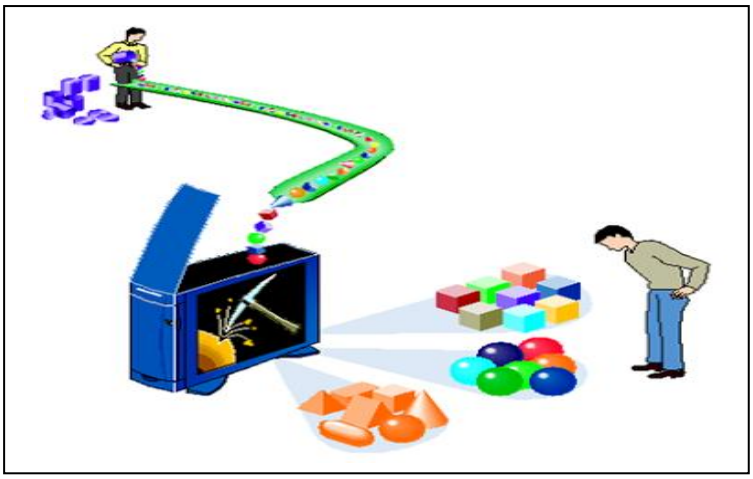
Fraud management is a knowledge-intensive activity. The main AI techniques used for fraud management include:

- Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.
- Expert systems to encode expertise for detecting fraud in the form of rules.
- Pattern recognition to detect approximate classes, clusters, or patterns of suspicious behaviour either automatically (unsupervised) or to match given inputs.
- Machine learning techniques to automatically identify characteristics of fraud.
- Neural networks that can learn suspicious patterns from samples and used later to detect them.

Other techniques such as link analysis, Bayesian networks, decision theory, land sequence matching are also used for fraud detection.

IV. DATA MINING

The word "Mining" refers to the extraction of valuable things like minerals from the earth. However, **data mining** is the process by which we can extract interesting patterns and knowledge from huge amounts of data. The data mining is a relatively new field of study and research and has generated huge interests among business communities. It is an important part of business intelligence which deals with how an organization uses analyses, manages and stores data it collects from various sources to make better decisions. Data mining has the answers to all these questions. Data mining can help organizations to have useful insights into its business from the data it has collected over the years and take better decisions.



A. Steps Involved in data mining

There are various steps that are involved in mining data.

1) Data Integration: First of all the data are collected and integrated from all the different sources.

2) *Data Selection:* We may not use all the data we have collected in the first step. So in this step we select only those data which we think useful for data mining.
3) *Data Cleaning:* The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get rid of such anomalies.
4) *Data Transformation:* The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.
5) *Data Mining:* Now we are ready to apply data mining techniques on the data to discover the interesting patterns. Techniques like clustering and association analysis are among the many different techniques used for data mining.
6) *Pattern Evaluation and Knowledge Presentation:* This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.
7) *Decisions / Use of Discovered Knowledge:* This step helps user to make use of the knowledge acquired to take better decisions.

B. Major tasks of Data mining

1) Summarization: Summarization is the generalization or abstraction of data. A set of relevant data is abstracted and summarized resulting in a smaller set which gives a general overview of data. For example, the long distance calls of customer can be summarized in to total minutes, total calls, total spending etc instead of detailed calls. Similarly the calls can be summarized in to local calls, STD calls, ISD calls etc.

*2) Clustering***:** Clustering is identifying similar groups from unstructured data. Clustering is the task of grouping a set of objects in such a way that object in same group are more similar to each other than to those in other groups. Once the clusters are decided, the objects are labelled their corresponding clusters, and common features of the objects in cluster are summarized to form a class description. For example, a bank may cluster its customer in to several groups based on the similarities of their income, age, sex, residence etc, and the command characteristics of the customers in a group can be used to describe that group of customers. This helps the bank to understand its customers better and thus provide customized services.

3) Classification: Classification is learning rules that can be applied to new data and will typically include following steps: pre-processing of data, designing modelling, learning/feature selection and validation /evaluation. Classification predicts categorical continuous valued functions. For example, we can make classification model to categorize bank loan application as either safe or risky. Classification is the derivation of model which determines the class of an object based on its attributes. A set of object is given as training set in which every object is represented by vector of attributes along with its class. By analysing the relationship between attributes and class of the objects in the training set, classification model can be constructed. Such classification model can be used to classify future objects and develop a better understanding of the classes of the objects in the data base. For example, from the set of loan borrowers (Name, Age, and Income) who serve as training set, a classification model can be built, which concludes bank loan application as either safe or risky.

4) Regression: Regression is finding function with minimal error to model data. It is statistical methodology that is most often used for numeric prediction. Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so cautions advisable [5] for example, correlation does not imply causation.

5) Association: Association is looking for relationship between variables or objects. It aims to extract interesting association, correlations or casual structures among the objects i.e. the appearance of another set of objects. The association rules can be useful for marketing, commodity management, advertising etc. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness and based on the concept of strong rules, introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {Onions, potatoes} {burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, Continuous production, and bioinformatics.

V. DATA MINING TECHNIQUES FOR FRAUD DETECTION

Data mining techniques can be classified into two categories

- *Supervised Learning for Fraud Detection:* This method uses supervised learning in which all the available records are classified as "fraudulent" and "non-fraudulent". Then machines are trained to identify records according to this classification. However, these methods are only capable of identifying frauds that has already occurred and about which the system has been trained.

- *Unsupervised Learning for Fraud Detection:* This method only identifies the likelihood of some records to be more fraudulent than others without statistical analysis assurance. Unsupervised learning is closer to the exploratory spirit of Data Mining as stressed in the definitions given above. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables.

A. Credit Card Fraud

Credit card fraud detection is the process of monitoring the behavior of the customers‟ transaction level through a period of time.

1) Types of Credit Card Fraud: The first type which is the most common is the application fraud. The individual will falsify an application to acquire a credit card. The individual will give false information about his/her financial status in order to receive a credit card.

The second type is assumed identity. Assuming someone's identity has been in the long-run form for credit card fraud. The individual will falsify a name with a temporary address.

The third type is financial fraud which happens when an individual wishes to gain more credit than he/she currently has. They will apply for a credit card under their own name, but the information regarding their financial status will be false.

The fourth is skimming technology. Magnetic card skimming is a small handheld device with the sole purpose of collecting and storing the information on any credit card.

The fifth type is never received issue. This type of credit card fraud involves the theft of the card while still in transit.

2) Data Mining Techniques for Credit Card Fraud: The first technique is the **Peer Group Analysis**. This type of analysis is an unsupervised method for monitoring customer behaviors over a period of time. For each individual that has a credit card account a "Peer Group" of accounts is created that exhibit similar behavior. As time goes by, the behavior of an account is tracked by those accounts in its peer group. If an account has subsequent behavior which deviates strongly from its peer group is thus considered to have behaved anomalously and is flagged as a potential fraudulent.

The second technique is the **Break-point Analysis**. This technique distinguishes spending activities supported from transaction information in a single account. Current transactions are matched up with prior spending activities to spot features, such as rapid spending and an increase in the level of spending, which would not essentially be captured by outlier detection.

B. Health Insurance Fraud

1) Health care Fraud: Health care fraud includes health insurance fraud, drug fraud, and medical fraud. Health insurance fraud occurs when a company or an individual defrauds an insurer. Frauds committed by a policyholder could consist of members that are not eligible, concealment of age, concealment of pre-existing diseases, failure to report any vital information, providing false information regarding self or any other family member, failure in disclosing previously settled or rejected claims, frauds in physician's prescriptions, false documents, false bills, exaggerated claims etc.

2) Data Mining Techniques for Health insurance Fraud: Data mining empowers a variety of insurance providers with the ability to predict which claims are fraudulent so they can effectively target their resources and recoup significant amounts of money. Data mining helps medical insurance company to focus, for example, on claims with high percentage of recoverable fraud, isolate factors which indicates a payment request has a high probability of fraudulence, develop rules to use them to flag only claims likely to be fraudulent, and ensure adjusters could review claims that are not only likely to be fraudulent but also have the greatest adjustment potential.

C. Telecommunication Fraud

1) Types of Telecommunication Fraud: Some common varieties of fraud in the telecommunications world.

Subscription fraud: Subscription fraud happens when someone signs up for service (e.g., a new phone, extra lines) with no intent to pay. In this case, all calls associated with the given fraudulent line are fraudulent but are consistent with the profile of the user.

- **Intrusion fraud**: This occurs when an existing, otherwise legitimate account, typically a business, is compromised in some way by an intruder, who subsequently makes or sells calls on this account. In contrast to subscription calls, the legitimate calls may be interspersed with fraudulent calls, calling for an anomaly detection algorithm.
- **Fraud based on loopholes in technology**: Consider voice mail systems as an example. Voice mail can be configured in such a way that calls can be made out of the voice mail system (e.g., to return a call after listening to a message), as a convenience for the user. However, if inadequate passwords are used to secure the mailboxes, it creates vulnerability. The fraudster looks for a way into a corporate voice mail system, compromises a mailbox (perhaps by guessing a weak password), and then uses the system to make outgoing calls. Legally, the owner of the voice mail system is liable for the fraudulent calls; after all, it is the owner that sets the security policy for the voice mail system.
- **Social engineering**: Instead of exploiting technological loopholes, social engineering exploits human interaction with the system. In this case the fraudster pretends to be someone he or she is not, such as the account holder, or a phone repair person, to access a customer's account.

- **Fraud based on new technology**: New technology, such as Voice over Internet Protocol (VoIP), enables international telephony at very low cost and allows users to carry their US-based phone number to other countries. Fraudsters realized that they could purchase the service at a low price and then resell it illegally at a higher price to consumers who were unaware of the new service, unable to get it themselves, or technologically unsophisticated. Detecting this requires monitoring and correlating telephony usage, IP traffic and ordering systems.

2) Data Mining Technique for Telecommunication Fraud:

Signature-based Methods: These methods are considered supervised/unsupervised hybrid. Signatures are simply telecommunication accounts summaries that are time driven and capture the behavior of a credit card or computer user including frequency of use, type if use, length of use and location of use. As Ferreira et. al define it , a signature is a "vector of feature" whose variables are obtained from the coded fields of a collection of Call Detail Records (CDRs). Each CDR is in turn a vector of features that can be discrete, such as the calling city, or continuous, such as the calling duration. Examples of those fields from the CDR include:

- Data of call
- Length of call
- Destination of call (example: 39-0382-506224)
- Time of call
- Origin of call (example: 973-360-8430)
- Payer of call (example: 973-360-8430)

Updating Signatures: Signatures must be updated continuously in an event driven manner to be able to recognize fraud as it happens. In general, there are two variations of signatures. The first type of signatures has a time-oriented processing in which users' actions are accumulated, kept and processed during a time unit for later analysis. This type is usually referred to as "summary". These summaries have a shorter time span such as an hour, half a day or a complete day. The second type of signatures uses action-oriented processing, which are simply called "signatures". This type of processing makes the direct comparison between current actions related to the CDR against the signature. In order to capture the different behaviours in different situations, the signature needs longer time window which could be a week, a month or even half a year. The former type of processing is less costly considering the processing requirements of massive volumes of data used in the signature.

D. Money Laundering Fraud

1) Money Laundering: Money laundering is defined as the process of concealing or disguising the proceeds of a crime or converting those proceeds into goods and services. It allows criminals to infuse money obtained illegally into the stream of commerce, thus corrupting financial institutions and the money supply. While many definitions for money laundering exist, it can be defined very simply as any knowing use of the proceeds of criminal activity.

Money laundering is usually associated with crimes that provide a financial gain. This includes, but is not limited to, bank fraud, insurance fraud, mortgage fraud, health care fraud, securities/commodities frauds, advanced fee schemes, high yield and prime bank note schemes, Ponzi schemes, government fraud, corporate and occupational frauds, cyber crimes, public corruption, drugs, organized crime, and the financing of terrorism. Money laundering differs from other types of criminal acts in that it is not a stand-alone crime. The laundering of funds is typically a secondary criminal act – typically without proceeds from an underlying crime, there can be no money laundering.

2) Data Mining Techniques for AML:

- **Rule-based Approach**: Harmeet Kaur Khanuja et al.[2] 2014, proposed a forensic methodology for private banks, which included the monitoring of transaction audit logs as per Reserve Bank of India (RBI) guidelines to mark the suspicious transactions if any and the DempsterShafer Theory of Evidence to generate reports. In paper by Rajput et al.[3] 2014, they proposed an ontology based system for detecting suspicious transactions based on a set of Semantic Web Rule Language(SWRL) rules and domain knowledge. Nida S. Khan et al.[6] 2013, presented a Bayesian network (BN)-based approach which assigns calculates the customer‟s transaction behaviour score based on transaction history and later on generates an alert if a significant difference is detected in customers historical transactional pattern and the current behaviour. Suvasini Panigrahi et al.[13] 2009, proposed a system for database intrusion detection which made use of rule-based approach along with belief combination, database history and Bayesian learning component to collect the evidences about the transactional behavior, mark the transaction as suspicious and raise alarm.
- **Clustering-based Approach**: Clustering is the process of grouping the data into classes so that objects within the same cluster have high similarity and objects within different clusters are very dissimilar. There are different clustering methods in the literature and they have been successfully exploited for scientific datasets, spatial datasets, business datasets, etc. In AML, clustering is normally used for grouping transactions/accounts into clusters based on their similarities. This technique helps in building patterns of suspicious sequence of transactions and detecting risk patterns of customers/account. One of the most challenges in clustering financial datasets is their size, this technique, for instance, should deal with millions of transactions during hundreds/thousands of time instances.

- **Classification-based Approach**: Stefan Axelsson et al.[8] 2012, analysed the implications of using machine learning techniques for money laundering detection in a data set consisting of synthetic financial transactions and aimed to detect anomalies inside a data set of mobile money financial transactions by using the classification techniques to group transactions as suspicious or non-suspicious. Xingqi Wang et al.[14] 2009, proposed a novel algorithm to detect money laundering using an improved minimum spanning tree clustering, an analysis of similarity measure and distance metric.

VI. CONCLUSION

Fraud remains a challenge for businesses and organizations in many fields. Data mining is an effective method for detecting various types of fraud including telecommunication, credit card and medical insurance fraud as well as detecting intrusion to computer systems. This research paper has presented only a selection of the various data mining techniques for fraud detection in different fields. Some of those techniques have persisted and proven to be successful, while others are in the process of development and enhancement to better apply to new fraudulent acts. After all, it is not the organization alone who suffers from the consequences of fraud, but all the individuals and stakeholders related to that organization will be victims. Therefore, organizations are entirely accountable for learning the best practices and choosing the best method that matches their needs in order to safeguard against fraud.

REFERENCES

[1] Outlier detection- Jiawei Han, Micheline Kamber, Jian Pei, "*Data Mining: Concepts and Techniques*".
[2] Data mining tasks and Techniques" www.investopedia.com
[3] Shashidhar, H.V. and S. Varadarajan, 2011. *Customer segmentation of bank based on data mining-security value based heuristic approach as a replacement to k-means segmentation.* Int. J. Comput. Appli., 19:13-18.
[4] Harmeet Kaur Khanuja, Dattatraya S. Adane, "Forensic Analysis for Monitoring Database Transactions", Springer, Computer and Information Science Volume 467, pp 201-210, 2014
[5] Pradnya Kanhere "*A Survey on Outlier Detection in Financial Transactions*" Int. J Computer Applications (0975 – 8887) Volume 108 – No 17, December 2014
[6] "*Data Mining for Fraud Detection: Toward an Improvement on Internal Control Systems*", http://www.researchgate.net/publication/241153108
[7] "*Data Mining Concepts and Techniques*", Jiawei Han and Micheline Kamber.
[8] Ciro Donalek Ay/Bi 199 – April 2011 "*Supervised and Unsupervised Learning*"
[10] S.D.Gheware, *Int. J. of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2014,"Data Mining: Task, Tools, Techniques and Applications*"
[11] "*Data-Mining-Techniques*", http://www.statsoft.com/Textbook/Data-Mining-Techniques
[12] "*Money Laundering*", https://www.fbi.gov/about-us/investigate/white_collar/money-laundering

