Study of Fraud Detection using Big Data Approach

Debdutta Pal and Supriyo Pal

Calcutta Institute of Engineering and Management, Kolkata - 700040, West Bengal, India; barmanroy.debdutta@gmail.com
Kolkata Municipal Corporation, Kolkata - 700013, West Bengal, India; pal.supriyo@gmail.com

Abstract

Fraud is increasing proportionally with the expansion of cutting edge technology and the e-globalization that cause loss of billions of dollar worldwide each year. In spite of having modern technology and worldwide superhighway communication we are failed to achieve our goal of secure e-globalization. To achieve our goal we need an efficient and effective fraud detection system. Fraud detection is a method of isolating illegal acts that are increasing worldwide. The aim of fraud detection system is to reveal the nature of fraudsters by applying appropriate methodology and specific domain knowledge. The amount of data produced in fraud detection growing large day by day. This cause difficulty to analyze huge amount of data that require more knowledge to gain. Today, in real world to create an efficient fraud detection system it is not enough to apply only data mining technique because data has become an indispensable part of every economy, industry, organization, business function and individual. The Big Data conceive unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation and measurement errors. This paper includes different types of fraud that we may face in our everyday life and how the big data can improve the acceptability of fraud detection system now a days.

Keywords: Big Data, Data Mining, Fraud Detection

1. Introduction

Van Vlasselaer has mention in[1] “Fraud is an uncommon, well-considered, imperceptibly concealed, time-evolving and often carefully organized crime which appears in many types of form”.

This definition includes some challenging features of fraudulent activities that must be focused during development of fraud detection system. Firstly fraud is defined as “Uncommon”, due to it’s ambiguous and unpredictable nature. Fraudsters behave as innocent, so that they are not noticed and they can do their job precisely in well-planned manner, this is why they are defined as “well-considered” and “imperceptibly concealed”. Fraud is termed as “time-evolving” because, fraudsters can adapt new methodology to fight against advanced fraud detection system. Another most important characteristic of

**Author for correspondence*

fraud is "carefully organized crime", because sometime fraudsters are not individual person rather a group of well-organized people involve in performing fraudulent activities in different aspects. In this situation it is very difficult to detect such offence because it seems to be very normal job. A final word in the definition of fraud indicates that "many types of form" in which fraud occur. Fraud can be classified into various ways depending on fraudster's activities or depending on domain of fraudulent activities. Whether it is the fraudulent activities or fraudster's behavior detection technique needs to be smart enough to identify the crimes. Some of the very popular fraud prone areas are insurance fraud, health care fraud, credit card fraud, fraud in social network, money laundering, and telecommunication fraud. Whatever be the area of fraudulent activities the detection system should handle huge amount of data for detecting fraud. This cannot be achieved by the traditional data base system. So, we need to jump from SQL to NO-SQL that is from traditional database system to Big Data. With Big data technology we enter in a new era of database technology which has yet to be revealed.

Big data analysis technology is used to correct repetitive errors that may occur during handling huge amount of data in medical sector. To prevent health insurance from attackers' big data analytic should use the technologies like anomaly detection, business rules, database searches, social network analysis and text mining. Data mining is the most appropriate approach for handling huge amount data in fraud detection.

Data mining is an analytic process which is designed to explore market or business related data in search of consistent patterns. Then this pattern is used to formalize the findings by applying the detected patterns to new subsets of data[2]. It is a method of non-trivial extraction of previously unknown and useful information about data[3]. It reveals interesting patterns and relationships hidden in a large volume of raw data.

Section 2 contains a detail study about big data that includes characteristics, issues and challenges regarding Big Data. In section 3 Data Mining techniques for Fraud detection are deployed with its objectives and difficulties. Big Data Mining in the domain of Fraud Detection is described in section 4. Lastly the paper is concluded in section 5.

2. Big Data

"Big Data" refers to huge volume of data that cannot be handled by traditional data base system. It appears in a concrete large size that hide any information it its massive volume, which cannot be explored by traditional data mining technique. New technologies are required to be developed to handle such kind of large volume of data.

Characteristics: The Big Data is known by its 3V's featured, where 3V's include[4-6,11]

- Volume: It refers to the massive data volume produced or manipulated by an organization that must be further manipulated in order to get useful information.
- Velocity: It refers to speed of data processing. Some organization performs such activities that need real time responses. Distributed and parallel processing algorithms become very important for this reason. Analyzing of day-by-day big data flows, millions of detailed record that must be scrutinized to get behavior pattern identification is necessary during fraud detection in health insurance.
- Variety: It refers to the various types of structured and unstructured data that are manipulated by an enterprise during day by day activities.

With these 3V's another important feature of Big Data has to be mentioned that is:

- Veracity: It refers to the information level of trust granted by business decision factors. This feature is one of the key parameter in fraud detection in health insurance.

Big Data Analysis:
A technique of analyzing large volume of data to reveal all hidden information that are important for any business decision making is known as Big Data Analysis. Big Data Analysis consider data acquisition and recording, information extraction and cleaning, aggregation and representation, query processing, data modeling ,data integration , data analysis and interpretation. To proceed through these phases some challenges have to be faced like[7]

- Heterogeneity and Incompleteness: The difficulties of Big Data analysis is the fusion of structured and unstructured data. Incomplete data may generate uncertainty during data analysis phase.
- Scale & Complexity: Increase in volume of complex data cause difficulties during analysis. For this reason some new tools have to be developed.
- Timeliness: Large volume of data increase the time of processing that may generate problem when decision has to be taken just in time.
- Security and Privacy: Big data analysis is done on huge volume of data. Data are collected from various sources that may supply malicious data which is harmful during data analysis.

3. Data Mining Technique

Data mining is a field of research that focuses to understand the unknown data pattern from a large set of data. There are different data mining techniques like classification, clustering, regression, association rule that are applied on huge volume of data set for analyzing the them and discover any unknown pattern from them for decision making.

Data Mining is associated with (a) supervised learning: It is based on training data of known fraud and legal cases.

(b) Unsupervised learning: This learning process is done with data that are not labeled to be fraud or legal. An example of unsupervised learning is Bedford's law[8,10].

There are four tasks to be performed to accomplish data mining

1. Classification: It is the process of grouping Data into predefined groups.
2. Clustering: It is a modified version of classification. In clustering data are grouped into classes that are not predefined. Clustering is an unsupervised classification.
3. Regression: It is a technique where a method is developed to model the data with minimum error.
4. Association Rule: This approach is used to discover the relationship among different data. It monitors frequency that the data set that appears simultaneously.

4. State of the Art Analysis of Fraud

In this section different category of fraudsters are described with their goal of interest.

In the Figure 1, it is observed that each organization is accompanied by two types of fraudsters. Fraudulent activities may be initialized by management of the organization or by non-management staff. This kind of activities is known as internal fraud. An organization when deal with huge number of external entities then there is a possibility of external fraudulent activities. The external fraudsters have three basic profiles: the average offender, criminal offender, and organized crime offender. Average offenders show dishonest behavior when there is chance, sudden influence, or when wretched from financial hardship. Most risky external fraudsters are either individual criminal offender or group of organized criminal offender. They often hide their own identities and theft other's identity for accomplish their task.

To identify fraudulent activities, it is important to detect the cause of crimes. Fraud Triangle is a pictorial representation that explain the reason why fraudsters born. There are three components that together lead to offensive activities either by an individual person or by a group.

In the Figure 2, it is observed that there are three factors that can make an innocent individual to fraudster. Sometimes economic or social pressure motivate a person to commit fraudulent activities. The pressure may come from family or from management. In another way if there is a chance do some illegal activities that may lead to financial benefit then also fraudulent activities can take place. A pathological laboratory can produce huge amount of bill for very silly pathological test, because it is known that patient's family will be worried about the disease, they will not argue for huge bill. Last one is rationalization, when somebody justify the fraudulent activities. Fraudsters think that they are innocent and due to certain circumstances they are bound to do criminal activities.

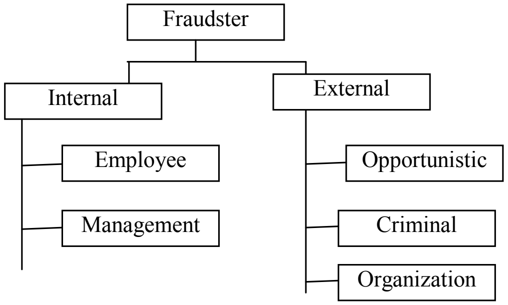

Figure 1. Categorization of Fraudster.

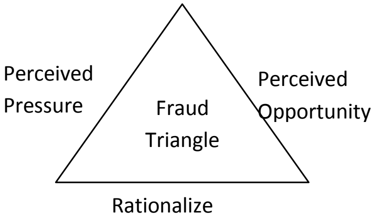

Figure 2. Fraud Triangle.

Depending on fraudulent activities two types of fraud are noticed[1]: They are:

1. Opportunistic Fraud: When a person takes advantage of inflating of a legitimate insurance claim. This is a very common type of fraud that may occur very frequently but the amount of money involve is limited.
2. Professional Fraud: This kind of fraud is performed by a group of fraud people who are very organized. Amount of money involved in this kind of fraud is very high. This kind of fraudulent activity is not so common.

5. Fraud Detection Technique

Depending on the approach of detection technique fraud detection can be classified into various types. Figure 3 depicts the classification of fraud detection system.

5.1 Expert Based Detection

Expert based fraud detection is one of the classic approach for detecting fraudulent activities. In this approach human intervention is most important. A person having experience and high end domain knowledge is appointed to monitor the fraudulent activities. Such criminal happenings have to be evaluated by the expert to judge how much it affect the system and find out a preventive measurement against such activities.

When the fraudulent activities are detected and confirmed, then two types of measures can be taken:

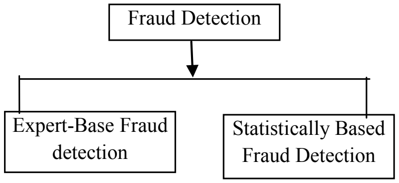
Fraud Detection
Expert-Base Fraud detection
Statistically Based Fraud Detection

Figure 3. Categorization of Fraud Detection Approach.

Corrective Measure: It includes retrospectively find out and address similar fraud cases that made use of same technique in the fraud detection.

Preventive Measure: It includes either an action of revealing the fraudster and exclude him from organization or incorporate new rules for detection of offensive activities in future.

Corrective measure is most effective way of reducing the effect of fraudulent activities.

Although the expert based fraud detection system is well accepted, but still it has some shortcomings

Like building of rule bases is very expensive as it needs the human expertise knowledge on specific domain. The rules should be kept up to date so that it can be applied to detect any criminal activities.

So a new methodology has been developed for fraud detection that is data driven method. There are some reason why researchers prefer data driven method over expert based system.

5.2 Statistically based Fraud Detection

The amount of data produced in fraud detection growing large day by day. This cause difficulty to analyze huge amount of data that require more knowledge to gain. Today, in real world to create an efficient fraud detection system it is not enough to apply only expert based detection technique because data has become an indispensable part of every economy, industry, organization, business function and individual. The Big Data conceive unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation and measurement errors. There are three reasons why data driven fraud detection is taking place of expert based fraud detection.

- Precision: This approach can uncover fraud pattern by analyzing huge volume of data that is difficult for a human expert.
- Operational Efficiency: This methodology allows real time decision making when a fraudulent activity is detected. For credit card fraud, the action has to be taken immediately, here data driven fraud detection is better approach.
- Cost efficiency: Data driven methodology is less expensive than expert based system in terms of economy and time.

5.3 Fraud Detection System

Following Figure 4 depicts how the fraud detection system performs its job. Fraudster and fraud monitor are actors where fraudster are performing fraudulent activities and fraud monitors are monitoring

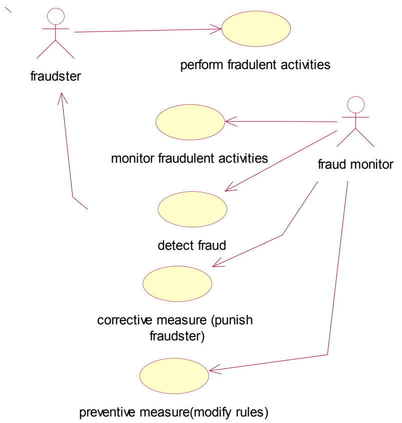
fraudster
perform fradulent activities
monitor fraudulent activities
fraud monitor
detect fraud
corrective measure (punish fraudster)
preventive measure(modify rules)

Figure 4. Fraud Detection Diagram.

At each layer of health insurance framework different kinds of data reproduced. Using Map Reducer tool of Big Data analysis reduce amount of data and generate a set of clean, relevant data. Then clustering approach of data mining technique is used to cluster the data according to their domain. Like the data related to TPA form a cluster then another set of data related to agent makes another cluster. Here Clustering is better because the policies to make cluster is not predefined. Depending on current rules and regulation the clustering parameter may differ. Then anomaly detection method is applied to identify the intruder. Another most promising approach for fraud detection in health insurance is Text mining. This technique is very efficient on big data volumes. Meaningful data are extracted and then analyzed by text mining algorithms to reveal abnormal or suspicious behavior of the intruder.

6. Conclusion

Few years back it was difficult and not a cost effective task for detecting fraud in health insurance. Big Data technology flourishes a new era in the fraud detection system. A Health insurance organization can now easily detect any fraud claim and take decision accordingly. But still now there is a limitation to apply big data technology in the domain of health insurance. To generate an efficient fraud detection system Big Data along with Data mining technique is much more acceptable. Big data analysis tool can be used to reduce the volume of data and generate clean, relevant data where s data mining technique is used to find out the fraud pattern and take decision using decision tree or regression approach. In future our aim is to create a hybrid fraud detection system using big data analysis and data mining technique.

7. References

1. Van Vlasselaer V, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B. GOTCHA! Network-based Fraud Detection for Social Security Fraud. Submitted to Management Science manuscript MS-14-00232, 2015.
2. Punde A, Daundkar K, Shelar S. A Review: Data Mining For Big Data. International Journal of Advanced Research in Computer Engineering and Technology (IJARCET). 2014 Oct; 3(10).
3. David JM, Balakrishnan K. Prediction of Learning Disabilities in School-Age Children Using SVM and Decision Tree. Int J of Computer Science and Information Technology. 2011; 2(2):829–35. ISSN0975-9646.
4. Halevi G, Moed H. The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature. Research Trends. Special Issue on Big Data. 2012; 30:3–6.
5. Ularu EG, Puican FC, Apostu A, Velicanu M. Perspectives on Big Data and Big Data Analytics. Database Systems Journal. 2012; 3(4):3–14.
6. Punde A, Daundkar K, Shelar S. A Review: Data Mining for Big Data. International Journal of Advanced Research in Computer Engineering and Technology (IJARCET). 2014 Oct; 3(10).
7. Jaseena KU, David JM. Issues, Challenges, and Solutions: Big Data Mining. Netcom, CSIT, GRAPH-HOC, SPTM – 2014. 2014; 131–40. © CS & IT-CSCP 2014.
8. Bolton R, Hand D. Statistical Fraud Detection: A Review. Statistical Science. 2002; 17(3): 235–55.
9. Bologa A-R, Bologa R, Florea A. Big Data and Specific Analysis Methods for Insurance Fraud Detection. Database System Journal. 2013; 4(4):30–9.
10. Phua C, Lee V, Smith K, Gayler R. A comprehensive survey of data mining-based fraud detection research. Xiv preprintar Xiv: 1009.6119, 2010.
11. Halevi G, Moed H. The evolution of big data as a research and scientific topic: overview of the literature. Research Trends. Special Issue on Big Data. 2012; 30:3–6.