International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 6

# A Survey on Methodology of Fraud Detection using Data Mining

**Kaithekuzhical Leena Kurien**
Research scholar, VTU-RRC, Bangalore, India

**Dr. Ajeet Chikkamannur**
Professor & Head, Department of Computer Science and Engineering, R.L. Jallappa Institute of Technology , Bangalore

## ABSTRACT

while many financiers use the Internet and social media to help them with investment decisions, these online tools can provide many benefits for investors and at the same time, same tools can make smart objectives for lawbreakers. These offenders are quick to adapt to new technologies – and Social media is no exception. Social media, such as Facebook, YouTube, Twitter, and LinkedIn, have become key tools for investors worldwide. Whether they are seeking study on particular stocks, background information on a broker-dealer or investment consultant, guidance on an overall investment strategy, up to date news or to simply want to discuss the markets with others, investors turn to social media. Social media also offers a number of features that criminals may find attractive. Fraudsters can use social media in their efforts to appear legitimate, to hide behind anonymity, and to reach many people at low cost.

***Keywords:*** *Fraud, Data Mining, Fraud Detection, Financial Fraud, Neural Network, Decision Tree*

## 1. INTRODUCTION

The definition of Data mining is the process of sorting through large data sets to recognize patterns and establish associations to solve problems through data analysis. Data mining tools allow organizations to predict future trends. The advantages of Data Mining are identifying patterns in the data and analyzing the behavior of the user. Data mining techniques are used in many research areas, including mathematics, healthcare, genetics, banking and marketing. While data mining techniques are a means to drive efficiencies and predict customer behavior, if used correctly, a business can set itself apart from its competition through the use of predictive analysis. Specific data mining benefits vary depending on the goal and the industry. Fraud can be defined as wrongful or criminal deception intended to result in financial or personal gain. Data mining is used in predicting, forecasting across many different application areas like health, credit card, sales, weather, satellite and finance. The financial industry companies use data mining tools to build risk models and detect fraud. Learning and observing fraud patterns from historical data can be used to combat fraud. It is impossible to be absolutely convinced about the legality and objective behind an application or transaction. Given the ground reality, the best cost effective decision is to tease out possible indications of fraud from the available data using mathematical algorithms.

Evolved from numerous research communities, especially those from developed countries, the analytical engine within these solutions and software are driven by artificial immune systems, artificial intelligence, auditing, database, distributed and parallel computing, econometrics, expert systems, fuzzy logic, genetic algorithms, machine learning, neural networks, pattern recognition, statistics, visualization and others. There are plenty of customized fraud detection solutions and software which protect businesses such as credit card, e-commerce, insurance, retail, health, banking, social media and telecommunications industries.

## 2. TYPES OF FRAUDSTERS

This section highlights the types of fraudsters and affected industries.

### 2.1 Fraudsters

Social media is landscape-shifting.[1]The use of social media by the financial services industry is rapidly accelerating .In growing numbers, the registered investment advisers are using social media to communicate with existing and potential clients, promote services, educate investors and recruit new employees. Firms use of social media must comply with various provisions of the federal securities laws, including, but not limited to, the anti-fraud provisions, compliance provisions and record keeping provisions. Hence fraud cases happening in e-commerce are increasing at an alarming rate.

The fraudster always stays alert and vigilant and has through knowledge about the industry h/she plans to target. Since years, each business is always susceptible to fraud internally or from external sources targeting the company/individual. In addition to internal and external audits for fraud control manually, data mining algorithms play an important role in analyzing the faulty transactions.

The fraudster can be an outdoor party, or parties. Also, the fraudster can either commit fraud in the form of a prospective/existing customer (consumer) or a prospective/existing supplier (provider). The external fraudster has three basic profiles: the average offender, criminal offender, and organized crime offender. Average offenders display random and/or occasional dishonest behavior when there is opportunity, sudden temptation, or when suffering from financial hardship.

In contrast, the more risky external fraudsters are individual criminal offenders and organized crime offenders (professional/career fraudsters) because they repeatedly disguise their true identities and/or evolve their *modus operandi* over time to approximate legal forms and to counter detection systems. Therefore, it is important to account for the strategic interaction, or moves and countermoves, between a fraud detection system's algorithms and the professional fraudsters' *modus operandi*. It is probable that internal and insurance fraud is more likely to be committed by average offenders; credit and telecommunications fraud is more vulnerable to professional fraudsters.

For many companies where they have interactions with many external parties, it is cost-prohibitive to manually check the majority of the external parties' identities and activities. So the riskiest ones determined through data mining output such as suspicion scores, rules, and visual anomalies will be investigated.

The main purpose of these detection systems is to identify general trends of suspicious/fraudulent applications and transactions. In the case of application fraud, these fraudsters apply for insurance entitlements using falsified information, and apply for credit and telecommunications products/services using non-existent identity information or someone else's identity information. In the case of transactional fraud, these fraudsters take over or add to the usage of an existing legitimate credit or telecommunications account.

There are other fraud detection domains. E-businesses and e-commerce on the Internet present a challenging data mining task because it blurs the boundaries between fraud detection systems and network intrusion detection systems.

[2] Shows how the financial advisors are using social media to promote or share investment advice .The retail customers are also using social media to streamline their investment portfolios giving enough room for fraudsters.

[3]Alibaba owned UCWeb browser is suspected of stealing data of Indian users. The alleged data theft is currently investigated by Hyderabad based Centre for Development of Advanced Computing (CDAC).

## 3. DATA AND DIMENSIONS

The attributes and data in the attributes when examined, the knowledge can be mined in terms of patterns or fraudulent transactions can be identified. The values in the attributes may be missing values, ambiguous data, repeated data values or outlier data which when examined gives potential insight for fraud detection.

### 3.1 Structured data

Structured data can be defined as the data with fixed width and size. The fields are discrete and the data in the fields can be accessed individually or combination of fields.

Generally, attributes can be binary, numerical

(interval or ratio scales), categorical (nominal or ordinal scales), multivalued attribute or derived attribute. Classification and Regression can be applied on supervised data.

### 3.2 Unstructured data

Unstructured data cannot predetermine the attributes and the data cannot be classified into fixed size or width. The unstructured data may be a mixture of different data types and the data would be difficult to classify.

## 4. METHODS AND TECHNIQUES

This section examines three major methods commonly used, and their corresponding techniques and algorithms.

### Overview

The various methods described below are observed and the advantages and disadvantages are discussed.

### 4.1 Supervised Approaches

The classification technique was found to be more effective in fraud detection. The Bayesian network is the technique used for classification task. Classification, given a set of predefined categorical classes, determines which of these classes a specific data belongs to. Decision trees are used to create descriptive models. Descriptive models are created to describe the characteristics of fault. The learner with the set of training data identifies a set of features (extracted from the set of transactions), to do fraud detection. After learning, he should be able to classify or identify a fraudulent from the set of transactions. As indicated in [4] the transaction which is identified as fraudulent as indicated with a risk score. Neural networks plays an important role and support vector machines are used for classification, regression and other tasks. [5] The applications like credit card fraud detection using Artificial Neural Networks [6] with three layers, the input layer, hidden layer and output layer. The set of features identified are applied to new transactions and the transactions can be classified as fraudulent or genuine. Neural Network based fraud detection system has been shown to provide substantial improvements in both accuracy and timeliness of fraud detection. The neural networks when applied on machines parallel speeds up the rule production for customer-specific credit card fraud detection [7]. Neural networks can process a large number of instances with tolerance to noisy data and has the ability to classify patterns on which they have not been trained. However, the disadvantages are it requires long training hours, extensive testing, retaining parameters like the number of hidden neurons, learning rate [8].The STAGE algorithm for Bayesian networks

[9] The Bayesian Belief Network in fraud detection and back propagation for Artificial Neural Network were used in the STAGE algorithm. STAGE repeatedly alternates between two stages of search: running the original search method on objective function, and running hill-climbing to optimize the value function. The result shows that Bayesian belief networks were much faster to train, but were slower when applied to new instances.

CART (Classification and Regression Tree) uses Gini index measure is used for selecting splitting attribute. Pruning is done on training data set. It can deal with both numeric and categorical attributes and can also handle missing attributes. Classification and regression tree provide automatic construction of new features within each node and for the binary target. [9]The ensemble bagging technique is used in classification and regression. It works by combining classifications of randomly generated training sets to form a final prediction. [10] Bagging uses decision trees. When a new instance has to be classified, each decision tree which is a weak learner votes for the instance and the process is applied to all the decision trees in the ensemble. The prediction is determined by maximum votes. This method is accurate and also stable. The paper stated in [11] utilizes naive Bayes, C4.5, CART, and RIPPER as base classifiers and stacking to combine them. The main issue is getting sufficient training data sets to determine fraudulent transaction.

The supervised learning should have good set of training data which would enable to identify the transactions are legitimate or fraud. The supervised approach usually has well defined data set and is much easier to achieve results as compared to unsupervised learning approach. But the unsupervised learning approach is usually more predominantly realistic and many areas of application can use unsupervised learning approach.

### 4.2 Unsupervised Approaches

Fraudulent operations mostly occur in a relatively small set of transactions from all the transactions. The

faulty transactions may be also being skewed or scattered across in the set of transactions. The noise factor that occurs from the data sets collected across all the areas for testing also makes it difficult to analyze fraud in unsupervised learning. Another difficulty that occurs is to analyze whether a particular transaction is fraud or genuine correctly. The data when examined may also contain some anomalies or outliers. Outliers are having data values significantly different or wider gaps as compared to other data in the data set.

In this paper [12] the two unsupervised algorithms of PCA (Principal Component Analysis) and SIMPLEKMEANS algorithm operating process and their reliability are discussed. The PCA has the ability to work on data sets independent of their content and sizes which can be large also and SIMPLEK MEANS clustering algorithm identifies whether the given transaction is fraudulent or not.

The Self Organizing Maps, an unsupervised learning technique produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. The result is clustering of input data.

[13].The transaction monitoring rules are automatically created and make possible their continuous improvement in an environment of dynamically changing information in an automated system.

In unsupervised learning approach, the application of credit card fraud contains increased possibility of skewed data [14]. The neural networks applied parallel can achieve faster results and the fraudulent transactions can be identified faster.

In Telecommunication fraud, the Agglomerative Hierarchical Clustering, an unsupervised learning algorithm constructs a tree like hierarchy, a dendrogram which contains all the values of 'k' clusters. [15] In agglomerative clustering the algorithm starts with each object representing a cluster and proceeds by combining other nearer objects into the single cluster. The dissimilarity measure is calculated and defined. The two distance measure used in the above paper was Euclidean measure and co relation. The paper also indicates the highest percentage of correct clusters occurs when correlation is the distance measure used. The Euclidean distance produced a distinct cluster of outliers regardless of class membership and correlation separated the normal cases from fraudulent cases.

In the unsupervised learning as shown in [16] the different tools available for statistical fraud detection are discussed and the areas where fraud detection technologies can be applied are also discussed.

In this paper [17], the unsupervised learning algorithm of clustering is applied to the dataset to find natural grouping in the data. The correct metric should be used for cluster analysis and the scaling, transformation and combining the variables to measure the distance between two observations.

The Peer Group Analysis as stated in [18] is a tool for monitoring behavior with respect to time in data mining situations. Each object is identified as target object and the behavior of the target object is compared to other objects. The target object showing behavior distinct as compared to other objects and the behavior of the target object is compared with the peer group.

### 4.3 Semi supervised Approaches

The semi supervised approach contains labeled data and unlabeled data. The labeled data may be less in number. [19] In this paper the comparison between unsupervised learning and semi supervised learning is performed using K Nearest Neighbor. In semi supervised approach, the negative instances are extracted and later apply fuzzy clustering for positive and negative example for outlier detection. The paper [20] discusses two step graph based semi supervised learning used in online auction .The social graph of online auction users and their transactions are analyzed using weighted degree centrality. This feature separates the fraudulent transactions and the legitimate transactions.

## CONCLUSION

Fraud cases are on increase with the onset of online financial transactions. The above paper explored the various possibilities of identifying or detecting of fraudulent transactions with legitimate transactions. The above survey has explored a wide array of all published fraud detection studies. It defines the different types of data where fraud can occur and the technical nature of data, performance metrics, and the methods and techniques. After identifying the

limitations in methods and techniques of fraud detection, this paper shows that this field can benefit from other related fields. Specifically, unsupervised approaches are more real time and occasionally the most difficult to find out whether the transaction or legitimate or fraudulent. The c approaches from intrusion and spam detection communities can contribute to future fraud detection research. However, the accuracy of .application from a set of transactions is still a problem in real world.

**REFERENCES**

Journal Papers:


[1] https://www.sec.gov/about/offices/ocie/riskalert-socialmedia.pdf

[2] www.investopedia.com/how-financial-advisors-are-leveraging-social-media.asp

[3]https://inc42.com/buzz/alibaba-ucweb-datatheft

[4] Detecting and preventing fraud with Data Analytics, ACL E-Book

[5] https://cloudtweaks.com/2014/09/use-supervised-unsupervised-data-mining/

[6] Sherman, E. (2002).Fighting Web Fraud. *Newsweek* June 10.

[7]Credit Card Fraud Detection using Bayesian and Neural Networks by Sam Maes, Karl,Bram[2002]

[8]Data Mining Techniques in Fraud Detection by Rekha Bhowmik ,*ADFSL Conference on Digital Forensics, Security and Law, 2008*

[9]Meta Learning Algorithms for Credit Card Fraud Detection Sanjay Kumar Sen, Prof. Dr Sujata Dash, 1Asst. Professor (Comp Sc & Engg.)

[10] Application of credit card fraud detection:Based on Bagging ensemble classifier by Masoumeh Zareapoor,Poourya Shamsolmoali.

[11] Chan, P., Fan, W., Prodromidis, A. & Stolfo, S. (1999). Distributed Data Mining in Credit Card Fraud Detection. *IEEE Intelligent Systems* 14: 67-74.

[12]Credit Card Fraud Detection with unsupervised algorithims by Maria R,Chloe O,Bignon G,Loic,Aristide P,France

[13]Machine Learning for Unsupervised Fraud Detection by Remi Dominques, INSA, KTH, Royal Institute of Technology2015,Sweden.

[14] Syeda, M., Zhang, Y. & Pan, Y. (2002). Parallel Granular Neural Networks for Fast Credit Card Fraud Detection. *Proc. of the 2002 IEEE International Conference on Fuzzy Systems*.

[15] Statistical Fraud Detection: A Review Richard J. Bolton and David J. Hand,2002

[16] An application of supervised and unsupervised learning approaches to telecommunications fraud detection by

Constantinos S. Hilas , Paris As. Mastorocostas Department of Informatics and Communications, Technological Educational Institute of Serres, Terma Magnisias, GR-62124 Serres, Greece.

[17] Unsupervised Profiling Methods for Fraud Detection by Richard J. Bolton and David J. Hand,Department of Mathematics

Imperial College,London,{r.bolton, d.j.hand}@ic.ac.uk

[18]Unsupervised Learning for Credit card Fraud detection by Prof Vikrant Agaskar,Megha Babariya,Shruthi Chandran,Namrata Giri,IJRET, Volume: 04 Issue: 03 | Mar -2017.

[19] International Journal of Innovative Research in Science, Engineering and Technology Copyright to IJIRSET Outlier Detection Using Unsupervised and Semi-Supervised Technique on High Dimensional Data Ms. Gayatri Attarde1, Prof. Aarti Deshpande2 M. E Student, Department of Computer Engineering, GHRCCEM, University of Pune, Pune, India 1 Professor, Department of Computer Engineering, GHRCCEM, University of Pune, Pune, India2.

[20] Two Step Graph-based Semi-supervised,Learning for Online Auction Fraud Detection, Phiradet Bangcharoensap1, Hayato Kobayashi2, Nobuyuki Shimizu2,Satoshi Yamauchi2, and Tsuyoshi Murata11 Tokyo Institute of Technology, Meguro, Tokyo 152-8552,Japan,phiradet.b@ai.cs.titech.ac.jp, murata@cs.titech.ac.jp2 Yahoo Japan Corporation, Minato, Tokyo 107-6211, Japanfhakobaya,nobushim,satyamaug@yahoo-corp.jp