Fraud Detection in Current Scenario, Sophistications and Directions: A Comprehensive Survey

M Kavitha
Research Scholar,
Bharathiar University,
Coimbatore.

Suriakala M.Sc., M.Phil., Ph.D
Assistant Professor,
Dept. of Computer Science,
Dr. Ambedkar Govt Arts College,
Chennai

ABSTRACT

Fraud Detection is one of the oldest areas of research. The requirement of an effective system that detects frauds effectively with zero loss exists until now. This is due to the increase in the technology, that influences both the ends; the user and the fraudster. Hence it becomes mandatory that the users need to stay a step ahead in this scenario. This paper discusses the changes that had taken place in the area of fraud detection. The flow of research from data mining approaches to machine learning approaches that were developed to defy the attacks are discussed here. It discusses the evolution of heuristic based mechanisms and graph based technologies that emerged in recent years. Further, it also discusses the need for Big Data based analysis in this domain. A few case studies are also discussed here to enable better understanding. Research challenges that exist in this domain in the current scenario are discussed along with the research directions.

Keywords

Fraud detection; Challenges; Graph databases; Graph based fraud detection; Big Data in fraud

1. INTRODUCTION

Increasing dependence on technology has led us to various advancements in E-Commerce and the M-Commerce domains. With the number of online users increasing exponentially and the types of services made available online increasing at a greater rate than ever before, serious threats in terms of security and privacy are also being dealt with. With the automation of several essential service sector elements like electricity bill payment, insurance, mobile recharge and telephone bill payments and also due to the emergence of E-Governance projects, almost every individual is forced to depend on online and mobile banking. This has led to the increasing number of Online Identity Thefts, Credit Card Fraud, Insurance Fraud, Banking Fraud, and Money Laundering for illegal activities, etc.

Another technological advancement made the process of identification and detection of these types of frauds possible in the first place. With the ability to collect, store and process huge amounts of data in real time, the potential risk involved in various scenarios were identified. The rate at which data is being generated today is way too high to be managed in the past. With such an advantage, more data is being collected than it was collected in the whole of the past. Data ranging from Server Logs, Clickstreams and Transactions to Data from Sensors, Cameras and other Ubiquitous Computing Devices which includes structured, semi-structured and unstructured data and the ability to handle them effectively using the current generation Big Data Platforms allows us to infer and reason every anomalous pattern which was not even close to a possibility in the past.

Even with such technological advancements, it is impossible to prevent fraud incidents all together. So our work concentrates on detection and prevention of fraudulent activities through advanced data analytics. The huge amount of data that enables us to detect these types of frauds in the first place, also possess several technological challenges that prevents us from utilizing the available data to the fullest possible extent. Challenges ranging from collection and storage of real time data to pre-processing, processing, visualization of processed data, etc. cripples our ability to identify fraudulent patterns to a greater extent.

Data imbalance is another major challenge preventing us from utilizing the data and the existing algorithms or techniques effectively. Imbalance in the data leads to skew and bias in results if not taken care of during the data modeling or pre-processing phase. Imbalance means that the distribution of data is not in the right proportion as in the example of credit card fraud detection process, the datasets will contain very few fraudulent transactions and millions of genuine transactions. The ratio of fraudulent transactions to genuine transactions is almost negligible that the fraudulent transactions may be eliminated as noise in the KDD process. So in-order to find the uncommonly common elements among the huge amount of noise algorithms that take into consideration the data imbalance is mandatory.

Also with the advent of Big Data, the imbalance is now on a multi-dimensional scale rather than on a single dimension as in the credit card example mentioned above. It becomes much harder when the regular or commonly common data is too huge and the fraudulent data is way too few. Now with big data the imbalance is much higher requiring much better solutions. The unstructured nature of the data makes it further complicated as traditional data mining techniques or algorithms require structured data.

This work considers the graph data model and the utilization of the current generation graph databases for anomaly detection. It allows us to model data on a single dimensional scale and then overlay them to obtain better results. For example, geographical data and temporal data can be modeled as different graphs and then overlaid to get better results. Also with the advent of the Apache Lucene Project, the unique feature of a search engines can be utilized. Search engines will not treat each word to be of same importance (as it is done in SQL), instead it assigns a frequency based index and frequently accessed elements will have a higher importance associated with it and hence will be on top. This helps us to better identify the uncommonly common entities resulting in effective anomaly detection at a faster rate.

2. LITERATURE REVIEW

The earliest fraud detection systems used data mining and statistics to extract fraudulent information from the available data. As the complexities of the frauds increase, these methods tend to get outdated. This lead to the emergence of advanced detection mechanisms and advanced architectures that facilitate these detections by improving the performance.

Explosion in the availability of data has directed us to a different scenario, an efficient mechanism that processes data containing the 3V's, the Big Data.

A discussion on whether a personalized model or an aggregated model suits the process of credit card fraud is presented in [5]. It discusses the pattern of working of aggregated models and the pattern of working of the personalized models. Personalized models similar to [6] and QTAM and RTAM aggregated models are used for the analysis. This analysis actually provides a very surprising conclusion stating that aggregated models works much better than the personalized models.

2.1.1. Data Mining based Fraud Detection Systems

A cost sensitive credit card fraud detection method that uses Bayes Minimum Risk classifier is presented in [10]. Even though many methods claim to include the actual cost of the detection process and the losses, they fail to do so in a practical environment. The authors of [10], claims to provide realistic views of the monetary gains and losses occurring due to fraud detection.

A method that concentrates on providing effective fraud detection using imbalanced data is presented in [15]. The basic working of any fraud detection algorithm is on imbalanced data. Most of the detection techniques ignores this aspect, which acts as a huge bottleneck when these systems are ported to the real environments. This method considers an extremely sparse and imbalanced data environment for performing the fraud detection process. Contrast Miner is used by [15] to mine the patterns that tend to point out fraudulent behavior. It uses three algorithms, contrast pattern mining, neural network and decision forest and integrates the results to present the final result.

A cost sensitive decision tree approach that tends to minimize the sum of misclassification costs in credit card fraud detection is presented in [13]. This method actually presents multiple cost sensitive decision tree approaches that provide better solutions when compared to other naïve approaches. Implementing these models in the real data sets, this method claims that it can be readily implemented in the real-world systems. It also claims that metrics such as TPR and FPR are not suitable for these classes of problems; hence it presents its own performance metric, which is the percentage of available usable limits saved.

A rule based fraud detection method that provided huge improvements in the detection process is described in [9]. This proposes to be a real time system that has been implemented in a Turkish Bank. it uses the MBO algorithm, that is based on the flight of birds. Improvements have been observed by comparing the current solution with the neighboring solutions and working accordingly. It claims to have improved the detection process and hence reduction in cost has been observed.

Usage of malware forensics in the area of fraud detections is explored in [16]. This method tends to detect frauds prior to occurrence by analyzing the anomaly prior patterns and the transaction patterns of normal users.

2.1.2. Heuristic based Outlier Detection

A transaction scoring method that uses genetic algorithm and scatter search to determine credit card frauds is presented in [14]. By scoring each transaction, this method claims to minimize the wrongly classified number of transactions. This method actually tries to improve the performance of an existing fraud detection system by tuning the parameters. It claims to have improved the performance by about 200%. Similar approaches providing fraud detection using Ant Colony Optimization and Particle Swarm Optimization have been discussed in[2,3,4].

2.1.3. Graph based Fraud Detection Systems

Mutual nearness based rank detection algorithm is proposed in [11]. It performs the process of anomaly detection by detecting the outliers. It provides an outlier score to all individual transactions. The transactions with outlier scores above the given threshold are considered to be outliers. It works on the basic principle that a centrally located object (here transaction) will tend to have many neighbors, or in other words similar transactions, hence it tends to have a very low outlier score, while a transaction at the periphery tends to have a high outlier score and hence has a high probability of being an outlier. The advantage of this approach is that it considers an average mutual nearness criteria, hence it can detect clusters of any shape effectively, unlike other mechanisms that tends to perform effectively only with certain shapes of clusters.

A scalable distance based outlier scheme that works on high volume data streams is presented in [12]. Unlike the usual scenarios that work on static data, this method takes the real time constraints into consideration, by processing on streaming data, which is the actual case of transactions generated by credit card transactions. Further batch processing has proven to be an outdated method when considering this area of fraud detection. The basic principle of this method is detecting distance based outliers. It uses the process of lightweight probing to determine the outliers. It also uses a principle of 'life-span aware prioritization' to leverage the temporal relationships among transactions to prioritize the order of processing of the data.

2.1.4. Big Data based Fraud Detection Systems

Evaluation of accuracy provided by the Hadoop MapReduce environment on the Credit Card Fraud detection data is presented in [7]. The Negative Selection algorithm is parallelized in the Hadoop environment for determining the accuracy. It uses the basic format of determining the outliers using the Euclidean distance and comparing it with the threshold. While the map function performs the distance calculation, the reduce function calculates the average distance and records the output. The map function also calculates the affinity threshold. Cost, False Negative Rate, detection rate and true positive rates were considered as the measuring criterion. The authors claim that the proposed system shows a considerable improvement in all the cases. A similar method that uses Euclidean distance to predict the outliers was proposed in [8]. It also utilizes the Hadoop environment to perform the operations.

3. CHALLENGES FACED BY A FRAUD DETECTION SYSTEM

3.1. Imbalanced Dataset and Incomplete Data

One of the major issues to be looked at in a fraud detection scenario is the Imbalanced nature of the dataset. In order to provide accurate results, the fraud detection applications require actual data rather than artificial data. The dataset providers, being bound to the confidential law, will not be able to provide the data as such. Data anonymization is carried out before making it public. Further, due to the real time nature of the data, it is bound to contain illegal values, missing or NULL values and inconsistent data [1]. Using the data as such will lead to misclassifications. Hence data should be cleaned prior to the usage. The three general phases include extraction, transformation and loading. But, in real the cleaning process should tailored according to the data being used.

3.2. Transaction Diversity

Another major challenge facing the fraud detection scenario is the customer himself. In general fraud detection mechanisms can be classified based on already existing fraud patterns (misuse based) or outliers from existing patterns (anomaly based). The drawback is that both these methods have their own downsides considering the human interaction involved. A normal transaction has all the chances of resembling an anomaly or a misuse based fraud. This proves to be huge downside when designing a fraud detection system. Due to the involvement of customers, it becomes crucial to design the system with least false positives. This leads to a compromise that sometimes costs a few actual fraudulent cases for the banks.

3.3. Real Scenario : Big Data

Even with the available real time data, developing a fraud detection system and deploying it in the real scenario proves to be a challenge, the reason being huge data availability and large data velocity, in short the Big Data. Due to the increase in usage of electronic money, the number of transaction inflow has increased to a large extent. Further, the number of record inflows per time unit has also shown a drastic increase. This results in an increase in the complexity of the detection system. Hence a fraud detection system that is to be developed for the current scenario should be capable of processing a large number of records in a short span providing accurate results is required.

3.4. Emerging New Patterns of Fraud

The technology boom in the last decade has not only seen an increase in the adoption of technology by masses, it has also seen an increase in the misuse of technology. As the technology advances and sophisticated techniques for detecting and preventing frauds emerge, the system fights back using advanced techniques for performing fraudulent activities, maintaining the equilibrium. Breaking this equilibrium has become one of the hardest processes, due to the emergence of new techniques and technologies. The initial detection mechanisms requires statistical or data mining techniques, while the current scenario demands further sophistications leading to machine learning and heuristic methods. Due to the real time nature of the problem, it also demands quicker results, which proves to be the biggest challenge.

3.5. Visualization

Algorithm based analysis have been in scope for quite some time. Even though this method provides effective results, certain scenarios cannot be understood without visualizing them. The lack of usage of visualization techniques in this scenario certainly handicaps the analysts from certain patterns.

3.6. Fear of False Positives

The biggest problem encountered in the process of fraud detection is misclassification. Though true negatives prove to be financially costly for the banks, false positives incur heavier damages. A false positive is one that diagnoses a legitimate transaction to be fraudulent. When response to this alert is triggered, it affects the customer directly, which incurs damage to the good will of the organization. Hence reduction of false positives seems to be the major concern for many organizations when dealing with fraud detection applications.

3.7. Need for Online Prediction

Most of the fraud detection mechanisms, even though they provide effective detection, they have very high latencies. This cannot be accepted when it comes to applications such as fraud detection. Faster the fraud is detected, the better. Every second that passes after the occurrence of a fraud proves to be an advantage to the fraudster. The major hurdle faced by these systems is the sheer velocity and volume in which these transactions are committed. Analyzing the transaction streams as they come and predicting accurate results proves to be a huge challenge for the detection models.

4. CASE STUDIES

The following presents a few case studies of some of the mostly encountered frauds.

4.1. Insurance Fraud

Insurance frauds are categorized to be one of the most occurring frauds. These can be faked easily with records hence it also remains to be one of the undetectable frauds. Due to the large number of claims in this area, verification also seems to be a problem hence most of these frauds go unnoticed, except for the ones involving a huge amount of money.

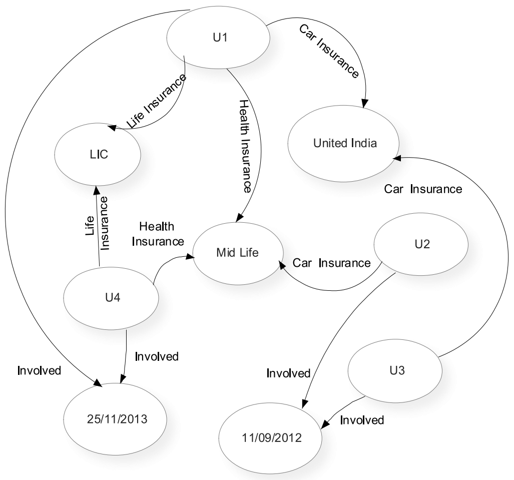


Fig 1: Sample Scenario (Insurance Fraud)

Figure 1 represents a sample scenario of three users and the insurance they have purchased. The dates represent their claims. Let us consider a sample scenario in which U2 and U3 claims on 11/09/2012 for their car insurance. By normal methods, it becomes difficult to visualize this scenario, since U2 and U3 have insurance in different organizations. Even if an accident has actually happened, it is obvious from the graph that either U2 or U3 will be able to claim, but here the claims seems to have been requested from both the users. Since their car insurances belong to different companies, this fraud goes undetectable.

4.2. Financial (Bank) Fraud

Financial fraud is mostly conducted by fraud groups rather than a single individual. Usually these are called fraud rings. In our scenario, let us assume that the bank requests for any two of the three identifications from a user before opening an account; an address, phone no or PAN card number. From the figure 2 representing the details of 4 account holders, it can be seen that multiple users have the same address or phone or PAN. But from the point of the bank, the repetitions cannot be identified. This is usually the format of operation of a fraud ring. They remain legitimate for a particular period, and then clears all the balance, takes up all the available loans and fills thecredit cards and finally swipes off the grid. This is called a Bust out. These varieties of frauds can be identified by finding the fraud rings.

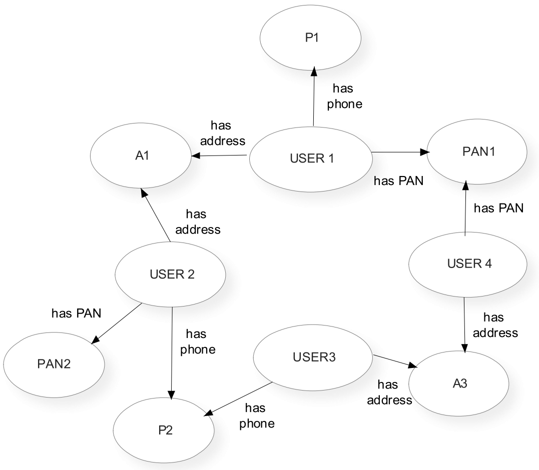


Fig 2: Sample Scenario (Financial Fraud)

4.3. Credit Card Fraud

This method of fraud is usually conducted by individuals who are not part of the card holders. This type of fraud has a typical difference from other frauds. The point of reporting of a fraud is usually the attack and the exploitation occurs at an earlier point in time. The figure 3 represents a transaction scenario for three users.

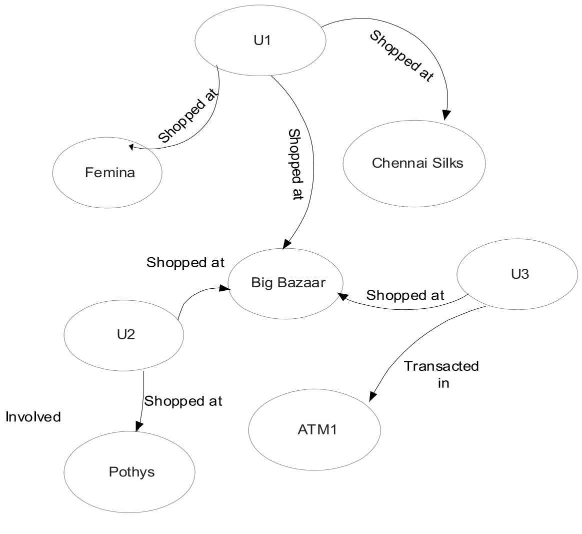


Fig 3: Sample Scenario (Credit Card Fraud)

If U1 and U2 reports an attack (a purchase in shop X that they have not made), then analyzing the point of victimization (X) is of no use. Instead, their transaction flow is to be analyzed.from the figure – it becomes clear that a common transaction point for U1 and U2 had been Big Bazaar. Hence a warning can be provided to U3 (who has also shopped in Big Bazaar) to modify their credentials [17].

5. RESEARCH DIRECTIONS

5.1. Personalized Vs Aggregated Models

General models built by any fraud detection system deals with performing operations on data considering the entire data as a single entity. Analysis of an aggregated data tends to provide a bird's eye view of the problem and a solution to the problem as a whole. But in a fraud detection scenario, the requirement is for a solution that predicts fraud for a single user's pattern. Hence personalized model that analyzes individual records and provides user level prediction seems to be a lucrative option. But experimentation shows that it is not true [5]. A model thatcoalesce individual predictionsalong with group predictions to provide appropriate results would be an option that can be explored with.

5.2. Finding Fraud Rings

Some frauds such as insurance are carried out by individuals, while others such as financial frauds are carried out usually by a group rather than a single individual. These schemes tend to be organized crimes that follow certain patterns. Identifying a common entity that tends to connect other entities which are supposed to be disjoint is one of the methods for identifying fraud rings. This method tends to find connections between entities that prove to be the point of identifying fraud. This process is usually carried out and works efficiently in graph databases rather than an RDBMS.

5.2.1. Finding Unusual Connections

Organized crimes usually tend to follow certain patterns or activity flows. Money laundering could be taken as a sample scenario for this case. In this case finding the fund flow from source to destination can provide a good overview of the money flow. The flow need not be necessarily direct. It can tend to move towards several intermediate entities before reaching the final destination. These structures can be best analyzed only when they are visualized.

5.3. GPU based Analysis

As discussed earlier, explosion in the number of transactions has lead to an increase in the number of transactions generation in a single time slot. Since the data generated tends to be individual transactions, the same process is to be carried out for each transaction. GPUs, can be used to perform these tasks in parallel, which tend to be the integral property of GPUs. A GeForce GTX 210, of NVidia contains thousands of cores and costs <2000, hence usage of CPU's in such tasks will not only prove to be efficient, but also cost effective.

5.4. Hadoop based Analysis

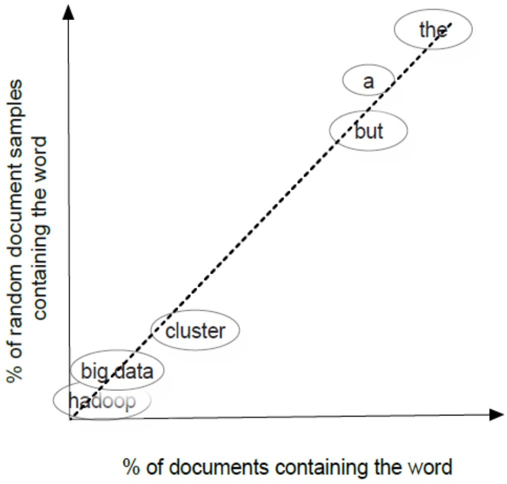
The usage of Hadoop has seen a huge rise in the recent years due to its parallel processing nature. Due to the huge volume of data associated with our current application, it acts as one of the best candidates of Hadoopable problems. Flexible accommodation of commodity hardware in Hadoop environment makes it one of the low cost and effective solutions for the problem of Fraud Detection.

5.5. Visualization

Though analysis using algorithms proves to provide effective solutions, there exist some hidden patterns which can be identified only with the help of human intelligence and analysis. Even the most complex of algorithms cannot replace human analogies. Hence a supervised visualization methodology using graphs can be developed, that can visualize the results to return patterns that were initially not decipherable using data based analysis. Uncovering structures in the graph dataset can lead to varied and much better solutions.

5.6. Search Engine based Model

One of the recently growing technologies is the concept of elastic search. This is a search engine based methodology that has its roots in Apache Lucene, an open source search engine. This works on the basic nature of a search engine that 'counts' everything provided to it. Frequency of occurrence of the transactions corresponding to a single individual can be mapped using the TF-IDF method to determine the outliers effectively. The search engine based analysis is performed by plotting the % of documents containing the word in x axis and the % of documents containing the word from a random sample of documents in the y axis. This usually forms a diagonal connecting the points (0,0) to (100,100). The words occupying the top right corner of the diagonal corresponds to mostly occurring cases and the ones occupying the bottom left corresponds to the rare occurrences. The nodes that occur in the region to the top left form the uncommonly common samples. This can be identified by overlaying the diagonal on a graph with x axis as the % of documents containing the word and the y axis as % of search results containing the word. This tends to shoot up the general uncommon words, but common in the current search. A sample graph is shown in Figure 4. This scenario can be directly mapped to the fraud detection system by considering the transactions and transaction sources. By the process of overlaying the transactions that are outliers in the system, or in other words dissatisfied transactions can be identified from the top left corner, which depicts the uncommonly common scenario.



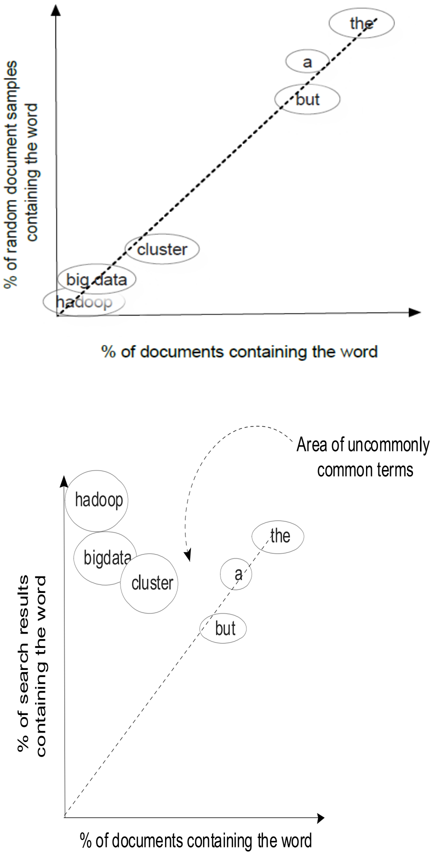


Fig 4: Sample graph depicting uncommonly common words

6. IMPLEMENTATION SUGGESTIONS

Due to massive data contained in the problem being analyzed, there are two viable options in order to determine efficient solutions. The first being GPU based processing and the second choice being Hadoop. The possible options available in this scenario are to port conventional algorithms to these architectures or to develop algorithms tailored to these environments. The most efficient and cost effective option, and the tradeoffs that are to be incurred by following each of these methods can be determined only by experimentation.

Further, as discussed in the above section, visualization can also provide various deep insights on hidden patterns. Various tools are available for processing a data set in the form of a graph. Neo4j [17], FlockDB, Allegro Graph, GraphDB and Infinite Graph are some of the frequently used graph databases. Each of the mentioned database software has its own advantages and the corresponding one can be used for effective problem management. Each of these datasets requires mastering of queries specific to them, hence portability is not an option while considering graph databases.

7. DATA SET DESCRIPTION

Due to the generic nature of the problem, a single dataset is not considered, instead, our research direction concentrates on developing a generic architecture that can be trained and utilized for all typesfraud detection process. Hence a variety of datasets have been used for analysis. Health care and insurance datasets have been analyzed with the CMS data set [18,19]. Enron email dataset is used for link analysis [20].

Credit card fraud detection is carried out using Australian and German credit datasets [22,23] and credit card dataset available in UCI repository [24].

8. CONCLUSION

Fraud detection methods that were developed from the early stages to the current technologies are categorized and analyzed. This paper provides their pros and cons and the current requirements for the fraud scenario existing in the present situation. Challenges faced while designing a fraud detection system are discussed in detail and the research directions are provided. Each of these directions leads to a different area of research and is disjoint from each other. Our future works will deal with providing solutions to the challenges posed by the environment utilizing the above mentioned directions. Further, analysis will be carried out to provide unified methods rather than separate methods for each of the sub-categories of frauds. Analysis will be carried out and the research will be directed towards creating a unified framework that trains with a specific category of fraud and provides effective detection and even prevention if possible.

9. REFERENCES

[1] Rahm, Erhard, and Hong Hai Do. 2000. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. 23.4: 3-13.

[2] Liu, Ou, et al. 2009. On an ant colony-based approach for business fraud detection. Emerging Intelligent Computing Technology and Applications. Springer Berlin Heidelberg, 1104-1111.

[3] Jia-jie, Shen, 2012. Electronic transaction fraud detection based on improved PSO algorithm. Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on. IEEE.

[4] Elías, Arturo, et al. 2011. Outlier analysis for plastic card fraud detection a hybridized and multi-objective approach. Hybrid Artificial Intelligent Systems. Springer Berlin Heidelberg, 1-9.

[5] Alowais, Mohammed Ibrahim, and Lay-Ki Soon. 2012. Credit Card Fraud Detection: Personalized or Aggregated Model. Mobile, Ubiquitous, and Intelligent Computing (MUSIC), 2012 Third FTRA International Conference on. IEEE.

[6] Rong-Chang Chen; Shu-Ting Luo; Xun Liang, Lee, V.C.S. 2005. Personalized Approach Based on SVM and ANN for Detecting Credit Card Fraud. Neural Networks and Brain, ICNN&B '05. International Conference on , vol.2,no., pp.810-815, 13-15.

[7] Hormozi, Elham, et al. 2013. Accuracy evaluation of a credit card fraud detection system on Hadoop MapReduce. Information and Knowledge Technology (IKT), 2013 5th Conference on. IEEE.

[8] Hormozi, Hadi, et al. 2013. Credit cards fraud detection by negative selection algorithm on hadoop (To reduce the training time). Information and Knowledge Technology (IKT), 2013 5th Conference on. IEEE.

[9] Duman, Ekrem, AyseBuyukkaya, and IlkerElikucuk. 2013. A Novel and Successful Credit Card Fraud Detection System Implemented in a Turkish Bank. Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. IEEE.

[10] Bahnsen, Alejandro Correa, et al. 2013. Cost sensitive credit card fraud detection using Bayes minimum risk. Machine Learning and Applications (ICMLA), 2013 12th International Conference on. Vol. 1. IEEE.

[11] Gurjar, Ram Niwas, Neeraj Sharma, and ManojWadhwa. 2014. Finding outliers using mutual nearness based ranks detection algorithm. Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on. IEEE.

[12] Cao, Lei, et al. 2014. Scalable distance-based outlier detection over high-volume data streams. Data Engineering (ICDE), 2014 IEEE 30th International Conference on. IEEE.

[13] Sahin, Yusuf, SerolBulkan, and EkremDuman, 2013. A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications 40.15: 5916-5923.

[14] Duman, Ekrem, and M. HamdiOzcelik. 2011. Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications 38.10: 13057-13063.

[15] Wei, Wei, et al. 2013. Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web 16.4: 449-475.

[16] Kim, Ae Chan, et al. 2014. Fraud and financial crime detection model using malware forensics. Multimedia Tools and Applications 68.2: 479-496.

[17] http://neo4j.com/

[18] http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient.html

[19] http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Outpatient.html

[20] https://www.cs.cmu.edu/~./enron/

[21] http://weka.8497.n7.nabble.com/file/n23121/credit_fruad.arff

[22] http://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29.

[23] http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29