Fraud Detection in near Real-time

Sapna Gupta

Submitted as part of the requirements for the degree
of MSc in Data Analytics
at the School of Computing,
National College of Ireland
Dublin, Ireland.

December 2015

Supervisor Adriana Hava Olariu

Abstract

Businesses are facing critical problems due to the growth of financial fraud. Banking, insurance and tax revenue industries are losing billions of dollars due to fraudulent transactions every year. Currently fraud detection systems chase fraudulent transactions which have already completed. Due to tremendous growth of web based applications, the volume of online transactions are growing enormously. Data mining techniques along with big data processing systems are the best solutions for detecting fraud on large volumes of data in real time. This paper presents a review of the literature found in the field of fraud detection using various data mining techniques. This review shows that supervised machine learning algorithms like Natural Network, Bayesian Network and Random Forest are extensively used in fraud detection applications. Less research was conducted in the detecting fraud in real time using more advance data mining techniques along with big data stream processing frameworks. There is a need of fraud detection applications which can help to stop fraud in near real time.

Contents

Chapter 1

Introduction

Fraud is criminal deliberate deception to secure illegal or illegitimate gain. Fraudster is a person who commits fraud; he can be internal or external person in the organisation. Survey done by Ngai et al. (2011) shows that finance related fraud can be mainly categorised into banking Fraud, insurance fraud, securities and commodities fraud and other related financial frauds. This study is more focused on banking industry fraud.

Banks spends millions of dollars every year on technologies to reduce fraud and retain customer confidence, but these spends does very little in protecting banks. As per Kumar (2015) majority of the techniques are dependent on data patterns and transaction signatures from a subset of historical transactions. Financial institutions use complex statistical models which are designed based on historical fraudulent transactions. These models are used to determine if transactions occurring in real time are fraudulent or not. No focus is laid on detecting first time fraud as there is no history or previous pattern of that kind of fraud. As this model is based on a subset of data, this model is not comprehensive enough. As a result financial institutions are always catching up and most of the times first time fraud never gets detected. The other concern is the ability of banks to continuously upgrade this model as more and more frauds take place. Banks have to strike a balance between detecting fraudulent transactions and any bad publicity it generates if a legitimate transaction is mistakenly identified as fraudulent. Any shortcomings of current model offer us an opportunity to create new models which are more accurate.

Jiang & Luo (2014) highlighted that the development of advance machine learning algorithms along with big data processing frameworks made data analytic techniques to work mush faster and to be more efficient than ever before. After this development it is easy to work on large dataset in real time and to find fraud with better accuracy.

In this study we are trying to find how data mining techniques together with real time big data processing frameworks are used to develop comprehensive fraud detection systems that are capable of detecting fraud in real time. Hence our research question is "How can we detect fraud in on-line banking transactions in near real-time using advance data mining techniques and data stream processing frameworks?"

In this paper 20 journal articles published between 2010 and 2015 which discussed different machine learning techniques to detect fraud have been reviewed. Section 2 discusses these journal articles based on the datasets used , data mining techniques used and machine learning algorithm used. Section 3 presents the project plan to design a fraud detection application using data stream processing framework and advance machine learning technique. Finally, section 4 is the conclusion of all the study performed. This section also highlights challenges and important points to be considered before starting the research project.

Chapter 2

Related Work

A comprehensive literature review was carried out in order to have a clear understanding of the chosen topic and various machine learning algorithms already been used in fraud detection applications. The goal of this literature review was to find out what other researchers have already done on this topic and to find gaps in the previous researches and the current requirement in fraud detection systems. We are also trying to find the best suitable machine learning algorithm for detecting fraud in near real time.

To find research articles, following search filters were used in Google Scholar and National College Library Summon search engines.

“Fraud Detection” AND (“Data Mining” OR “Machine Learning” OR Stream Processing) = 6772 results (full-text/scholarly);

(“Banking Fraud”) AND (“Real Time” OR “Machine Learning”) in the abstract = 128 results (full-text/scholarly);

“Banking Fraud” in the abstract “Data Mining” in All Fields = 32 results (full-text/scholarly);

Research articles were obtained from ACM Library, IEEE Transactions, Science Direct and Springer Link Journals.

After reviewing the abstract, introduction and conclusion, only 20 articles were found relevant to the research question. These research articles are discusses under 3 subsections, 2.1 datasets used, 2.2 data mining techniques used by different researchers and 2.3 machine learning algorithms used.

2.1 Datasets used

Fig. 2.1 shows data processing steps carried out in a any fraud detection application. This figure also lists some of the commonly used data mining techniques (Sharma & Kumar Panigrahi (2012)). It shows after acquiring the data from different legacy systems, data is cleaned and integrated. Data is transformed before applying any data mining technique on it. Different data mining techniques and machine learning algorithms are applied on the transformed data to build a comprehensive predictive model. This Model helps to find patterns and trends of past fraud behaviour. Model developed is evaluated based on the false positive error rate. To reduce error rate a careful selection of data mining technique and machine learning algorithm is required.

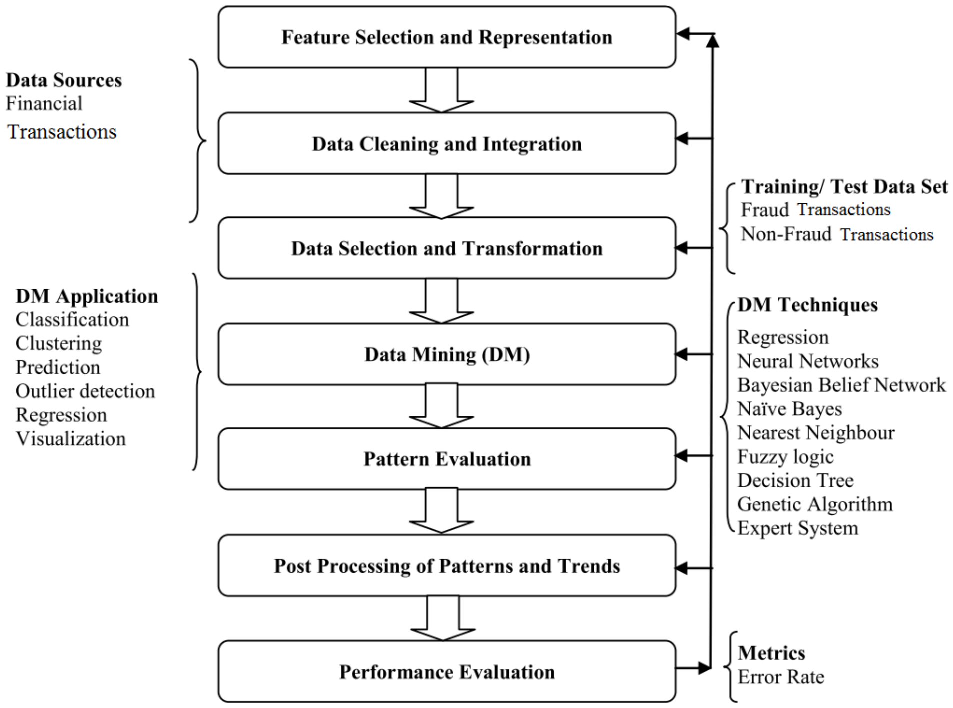


Figure 2.1: Data processing steps along with the list of available DM techniques.

After reviewing selected articles, it was found that credit card bank fraud is very well researched topic. Dataset plays a major role in the whole research project. It was found from these studies that credit card datasets are very highly skewed and have imbalanced classes. Fraudulent transaction are very less in number as compared to legitimate transactions. Fraudulent transactions ratio in the dataset plays a big role in the model accuracy therefore not every data mining technique work well with these datasets.

Bhattacharyya et al. (2011) used 50 million credit card transactions of 1 million credit cards of 13 months' data. Out of which only 2420 transactions are fraud. Data is highly skewed , therefore ,four training data sets are created (which have 15%,10%,5% and 2% fraudulent cases) using under-sampling. A test dataset is also created having 0.5% of fraud transactions.

Carminati et al. (2015) used a dataset of transactions from a large national bank, collected between December 2012 and August 2013.The data contains customer transactions related to Bank transfers 718,927, Prepaid Cards 100,688, Phone Recharges 71,362.

Bahnsen et al. (2013) used 80,000,000 credit card transactions with 27 attributes. This dataset contained only 3500 fraudulent transactions with fraud ratio of 0.467%. Research was carried on five different samples S1, S5, S10, S20 and S50, each one having a different percentage of frauds 1%, 5%, 10%, 20% and 50%, respectively.

Mahmoudi & Duman (2015) acquired dataset from anonymous bank in Turkey with fraudulent transactions ratio of about 10%. Research was carried out on a sample dataset with 8448 legitimate and 939 fraudulent transactions with 102 attributes.

Carneiro et al. (2015) used 645,538 real Internet credit card transactions with 37,359 fraudulent transactions. 80% of whole data was used for training, 10% for testing and 10% validation. There was no details found about the fraudulent transaction ratio.

2.2 Data mining techniques used

"Data mining is the computational process of discovering patterns in large data sets ("big data") involving methods like artificial intelligence, machine learning, statistics, and database systems" (SIGKDD (2014)). Data mining can be supervised or unsupervised. In supervised fraud detection mining, a new transaction is classified as fraudulent or legitimate based on the small set of fraudulent and legitimate transaction drawn from the historical transactions. According to Bhattacharyya et al. (2011) advantages of supervised model building is easy to implement but has some challenges. Main challenge is skewed class size of legitimate and fraudulent transactions, as model developed from this dataset does not have a very good accuracy for prediction. As per Rawte & Anuradha (2015) models based on supervised learning algorithms cannot find new patterns of fraudulent transactions and great amount of effort is required to train the model from the new fraudulent transactions and to deploy in the real world environment.

Unsupervised learning does not use previous knowledge or patterns to classify new transaction as legitimate or fraudulent. Chauhan & Shukla (2015) stated unsupervised learning is based on unusual behaviour and works very well with data streams. Unsupervised learning can help to detect any new or old unusual behaviours or patterns, whereas supervised learning can detect only pre-defined patterns. Unsupervised learning has the advantage of detecting any unusual behaviours which do not conform to normal behaviour and these patterns have lack of direction. So unsupervised learning can find any type of patterns. Un-Supervised learning has an inherent weakness of lack of direction and there may be instances when no new pattern was discovered in the sample dataset selected for the training. Rawte & Anuradha (2015) presented a study about health insurance industry and how to handle fraudulent claims. Claims are grouped according to type of disease and multiple claims are grouped this technique is called evolving clustering method. There are 2 types of detection systems used in the research project, one is supervised and other is un-supervised. Supervised detection cannot classify disease claims and un-supervised cannot detect out liars.

Moro, Cortez and Rita (2015) have outlined that have outlined that data mining techniques are further classified into different types i.e. classification, clustering, regression, prediction , outlier detection and visualisation (Moro et al. (2015)). Figure 2.2 is the graphical representation of these techniques along with list of commonly used machine learning algorithms (Sharma & Kumar Panigrahi (2012)).

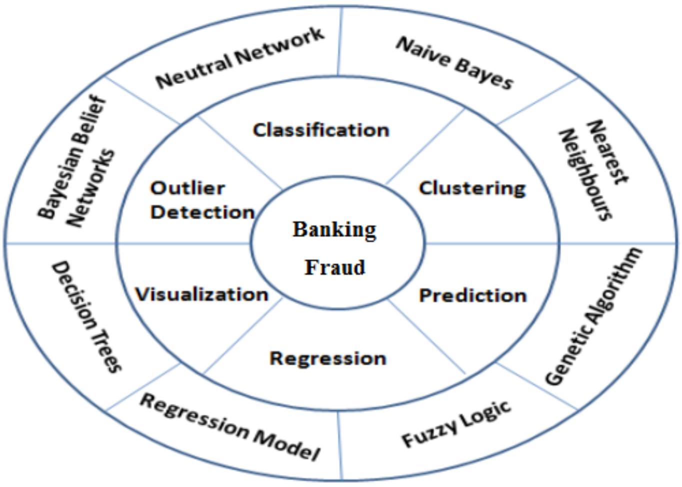


Figure 2.2: Conceptual model of data mining techniques applied on banking Fraud.

Classification is very widely used mining technique in fraud detection applications (Ngai et al. (2011)). Classification is supervised mining technique, consists of predicting some output based on the given input (Kirkos et al. (2007)). Classification algorithm process a training set of data and build a model based on the relationship found between the attributes and the outcome.This model is then tested on the test dataset, after testing the accuracy of the model is used to predict the outcome of the new transactions."Common classification techniques include neural networks, the Nave Bayes technique, decision trees and support vector machines SVM" (Liu et al. (2015); Kirkos et al. (2007); Zhou & Kapoor (2011)). SVM is a classification technique, using this technique hipper plane called decision boundary is drawn between classes of legitimate and fraudulent transactions. Whenever a new transaction comes into the system it is compared with the decision boundary (hyperplane) and is placed into either legitimate or fraudulent class. Bhattacharyya et al. (2011) proved random forest classification technique outperformed support vector machines and logistic regression techniques.

Clustering is unsupervised data mining technique, used to partition the given input dataset into unknown meaningful clusters or groups (Sorin 2012). Characteristics of transactions in one cluster are very similar to characteristics of other transactions in the same cluster, but are very different from the characteristics of transactions in other clusters. Carneiro et al. (2015) Suggest that transactions in each cluster should have very high intra cluster similarity and very low inter cluster similarity. As per various studies most commonly used clustering algorithm used for fraud detection are K-nearest neighbour KNN, K-mean and Nave Baye (Zhou & Kapoor (2011)). K-mean clustering algorithm is widely used and is very easy to implement. It requires a parameter k as input and divide transaction dataset into k clusters such that inter cluster similarity is very high and intra cluster similarity is very low. As this algorithm pre-defines the number of clusters, it becomes difficult to cluster new transactions which are not similar to any of the clusters present before (Ngai et al. (2011)).

Prediction is another type of data mining technique which is used to predict numeric ordered prospect values based on the patterns discovered from the transactional dataset (Seemakurthi et al. (2015)). Prediction modelling works well with continues ordered variable and is incapable to work with unordered categorical and discrete attributes in the dataset. Neural networks and logistic model prediction are mostly used prediction techniques (Ngai et al. (2011)).

Regression is a statistical data mining approach used to find the co-relation between one or more continuous independent variables and a continuous dependent variable (Ravisankar et al. (2011)). Linear regression and logistic regression are two most commonly used regression techniques in the field of data mining. Logistic regression work very well with binary outcome for example to find a given transaction is fraudulent or legitimate and is widely used in Fraud detection applications.Logistic regression does not perform well when there is imbalance in the dataset, like in fraud dataset where number of fraudulent transaction are very low as compared to legitimate transactions (Bhattacharyya et al. (2011)).

Outlier/anomaly Detection is very useful data mining technique to find fraudulent transaction in real time data streams Chauhan & Shukla (2015). Transactions which have very different behaviour than rest of the transactions are called outliers and these outliers require further analysis. The issue of anomaly detection is one of the most famous topics in data mining. Chauhan & Shukla (2015) proposed a two types of density based and partition based unsupervised K- means clustering algorithms to meet the challenge of volume and velocity of data streams. Density based fraud detection framework uses clustering technique to find outliers and partition based application uses euclidean distance to find outliers. Jiang & Luo (2014) used a "Field Programmable Gate Array (FPGA) -based intrusion detection system (IDS)", to process very large datasets.

Visualisation is easily understandable presentation of dataset (Ngai et al. (2011); Phua et al. (2010)). Every data mining application is not complete without any visualization. There are high chances of finding undiscovered patterns and relationship in the dataset, which were not discovered by other data mining techniques and can be easily discovered by visualization techniques. There are lots of advance visualization tools available, which uses colour, position, size and other visual characteristics to represent complex patterns found in the dataset. Most commonly used visualizations tools are Tableau, R-Plots and Rapidminer.

2.3 Machine learning algorithms used

Machine Learning is designing of algorithms which helps computers to learn from any given data without any human intervention. Data Mining is process of knowledge discovery from databases using machine learning algorithms. There are many machine learning algorithms used in fraud detection, which are discussed in this section.

Machine learning algorithm Artificial Immune System (AIS), as the name suggest imitates function of human body immune system. It detects the fraud the fraud as soon as it enters the system and learns the patterns of fraud, it gets simulated only by fraudulent transactions , once detector is simulated it again learn pattern of the new fraud. Halvaiee, N.S & Akbari (2014) used AIS (Artificial Immune System), and proposes a new model called AIS-based Fraud Detection Model (AFDM). The goal of this study is to increase the precision, decrease the cost and system training time. AFDM not only classifies the transactions but also allocated a risk factor to tell how risky the transaction is. To analyse performance of training model, five different training setup of cloud based Apache Hadoop framework were used, with different size of RAM , no of nodes , no of map functions and computing time was measured. Tests were done in serial and parallel mode and performance was compared. Measurement was based on computing time and Memory Cell Generation time. How performance varied by increasing number of CPUs with no of map functions being same tells us if number of nodes impacts performance and throughput. If no of nodes remain same, but number of map functions change, how does this affect the throughput and computing time. All these permutation combinations of various resources like CPU, RAM, no of nodes, no of map functions give a fair idea of which combination is suited in which scenario for best results.

It was proved by Bhattacharyya et al. (2011) that sensitivity, G-mean and weighted accuracy decreased with proportions of fraud in training data. Precision and specificity show an opposite trend on F-measure and accuracy. Logistics regression remains more or less same with various fraud data sets, whereas RF and SVM have decreasing on decreasing accuracy and increasing trend on F-measure. Random forest proved to be the most accurate out of all models. Bahnsen et al. (2013) added to the work done by Bhattacharyya et al. (2011) by also computing cost of false negative , gives rise to much better fraud detection results in the sense of higher savings. In this model also, RF outperforms then other algorithms (logistic regression (LR), C4.5) in terms of cost and F1-Score.

Carneiro et al. (2015) demonstrated how to normalize data for artificial neural networks by clustering of credit card transactions based on their attributes. Various attributes are clubbed into buckets using cluster analysis. Care is taken that there is no loss of data during this process of normalisation. Data normalisation has not been discussed by all other studies. Normalisation helps to quickly find patterns or anomalies and that can lead to finding fraudulent transactions rather quickly. It was rather surprising that data normalisation has not been used by any of them.

A most recent study Mahmoudi & Duman (2015) proposed Linear Discriminant Analysis (LDA) based on support vactor machine learning algorithm.Model is build by dividing the transaction into 2 regions whose boundaries were called decision boundaries the model tries to find a one dimensional hyperplane by which inter class similarity is minimised and inra class similarity is maximised. This model always try to reduce overlapping of the classes and maximise between class variance.

Carminati et al. (2015) used Principal Component Analysis (PCA), agglomerative hierarchical clustering algorithm which led to a satisfactory outcome. After testing the euclidean distance, they switched to the mahalanobis distance approach. DBSCAN was used for global profiling .This research was based on in-depth analysis of data set and identification of main features of data, which addresses the concerns like scarcity of training data and their extreme statistical imbalance. This model was refined by doing some real world fraudulent transactions. This model is highly proficient to detect on-line banking frauds in a semi supervised and unsupervised approach. Model developed by Sarno et al. (2015) was based on applying a set of positive association rules and then a negative set of association rules to finalise a list of fraudulent cases. The performance of the model build is relatively better than the model developed by other machine learning algorithms. The association rules capture fraud with given minimum confidence. If the minimum confidence is set high, the filters become too tight to catch the fraud. If rules are set as too low, many legal cases get termed as fraud.

Another important finding was that fraud detection systems using different machine learning algorithm can be compared using performance measures like accuracy, precision, specificity and sensitivity (Raj & Portia (2011);Bhattacharyya et al. (2011)).

These performance measures are based on four factors: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

How accurately negative (non-fraud) cases were predicted is termed as specificity.

$$(Specificity = TN/(FP + TN)). \tag{2.1}$$

How accurately positive (fraud) cases were predicted is termed as sensitivity.

$$(Sensitivity = TP/(TP + FN)). \tag{2.2}$$

Accuracy of the model is how accurately both positive (fraud) and negative (non-fraud) cases were identified.

$$(Accuracy = (TP + TN)/(TP + TN + FP + FN)). \tag{2.3}$$

How accurately the fraud cases were predicted is called precision.

$$(Precision = TP/(TP + FP)). \tag{2.4}$$

Furthermore, detailed review of the articles found is discussed under the following 2 subsections, dataset used and methodology used.

After reviewing all the studies found in this area , it looks like outlier detection using clustering algorithm works well with large volume of data, but it is still not clear which algorithm perform best in terms of time, as there was only one study (Halvaiee, N.S & Akbari (2014)) found, which compared time required to produce results in near real time.

Chapter 3

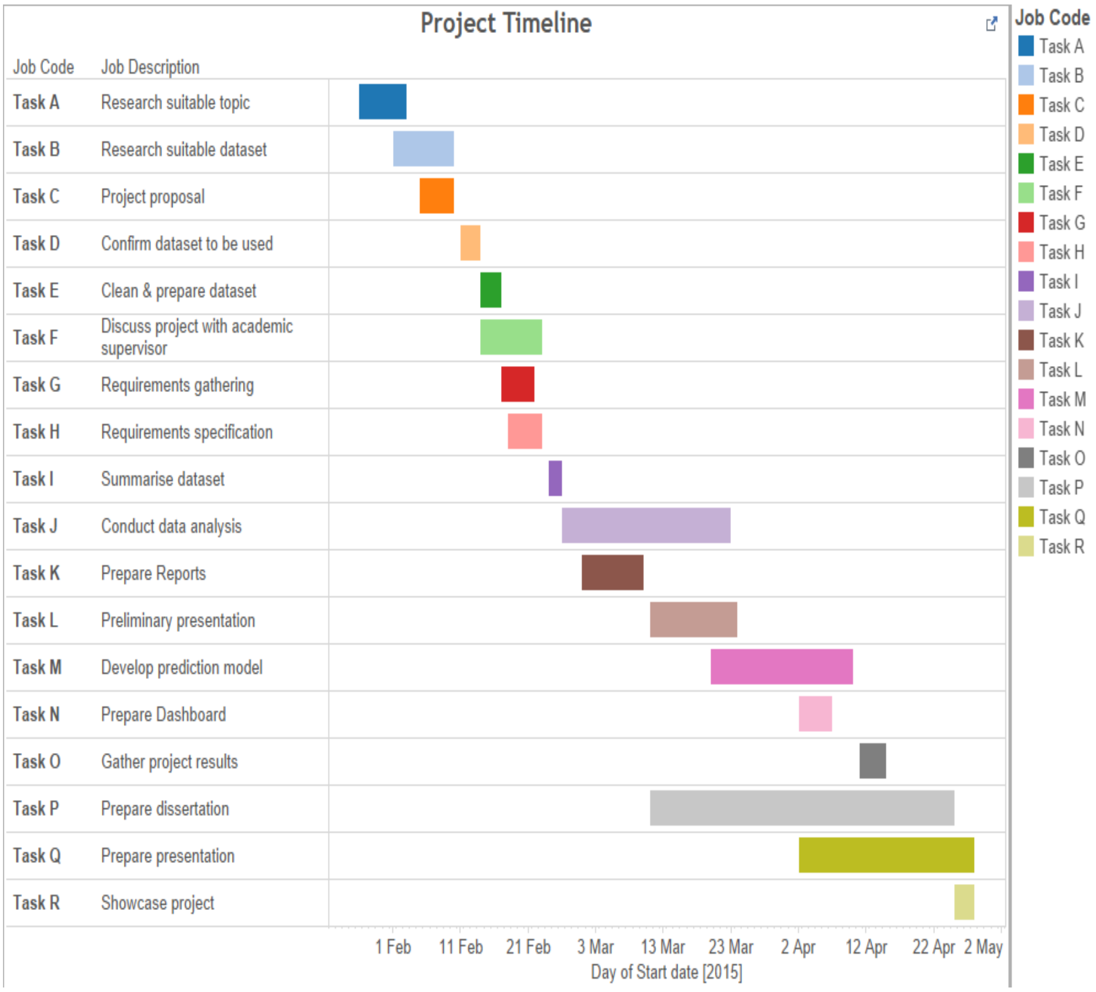
Project Plan

Data analysis and predictive analytics today are driven by large scale distributed deployments of complex pipelines, guiding data cleaning, model training and evaluation. A wide range of systems and tools provide the basic abstractions for building such complex pipelines for offline data processing, however, there is an increasing demand for providing support for incremental models over unbounded streaming data. Fraud detection is a very important application in financial domain. Goal of this research project is to build an efficient data pipelines using advance data stream processing systems like Apache spark Streaming or Apache Storm and design fraud detection models using advance machine learning algorithms like Hidden Markov Model (HMM). In this work, we focus on the problem of modelling such a pipeline framework and providing algorithms that build on top of basic abstractions, fundamental to stream processing.

Any data points detected as fraudulent is only a prediction. There is manual intervention required to verify if the predicted transaction is fraudulent and make the final call. Fraud detection is a use case for outlier detection in machine learning. Outlier is data point that is unlike other data points in whole dataset. There are many machine learning algorithms used for detecting outliers. The most widely used outlier detection algorithm used is HMM (Agrawal et al. (2015)). Many of the algorithms discussed in this literature review are not applicable for real time prediction. Real time fraud detection is only possible for those algorithms which can work together with data streams. HM model is build offline using historical data stored in Hadoop distributed file system. This model can be used in real time for incoming data streams using Apache Storm or Spark streaming framework.

3.1 Project Milestones

The project execution steps and time lines is shown in the Gantt chart below:



The first step will be to find suitable dataset and there is possibility of generating synthetic dataset for this research project. After acquiring the next step is to store acquired dataset into appropriate database, choice of database will depend on format and size of the dataset. As there is wide verity of open source tools and technologies available to choose from, it is very important to choose right set of tools to design the right pipeline for cleaning, processing and analysing of the acquired dataset to meet the aims of the project.

As found by Bhattacharyya et al. (2011) on-line transaction of a financial organisation is a dataset which is generally a combination of numerical and categorical attributes.

Transaction amount is numerical and customer details are categorical. Some of the categorical attributes have hundreds and thousands of classes. There can be several irrelevant and noisy attributes in the data. Exploratory analysis of the acquired dataset will be carried out. A careful data cleaning will be done to remove noisy and irrelevant attributes from the dataset. Data cleaning will be carried out before building of machine learning model, to produce more accurate results. Different data mining techniques will be applied on the dataset. Applied techniques will be compare in terms of performance and accuracy. Next step is to choose the best predictive model which can detect fraud in near real time. Results will be presented in the form of reports and dashboard. After gathering all the results a thesis documentation will be prepared. This task is one of the longest task in the whole project. Finally a brief presentation document is required for presenting the research work in front of the college faculty.

Chapter 4

Conclusions

Research shows that supervised and unsupervised machine learning algorithms are widely used in fraud detection. Supervised machine learning algorithms are easy to implement but has some challenges. Main challenge is skewed class size of legitimate and fraudulent transactions. Second challenge is that models based on supervised learning algorithms cannot find new patterns of fraudulent transactions (Bahnsen et al. (2013); Bhattacharyya et al. (2011)). Un-Supervised learning has an inherent weakness of lack of direction and there may be instances when no new pattern was discovered in the features selected for the training. There are some advantages and disadvantages of each of these machine leasing techniques, so there is a hybrid approach proposed. Most commonly used supervised machine learning algorithms are supervised support vector, decision trees, and random forest.

As per the survey done, fraud detection systems can be compared with one another based on the accuracy, precision, specificity , sensitivity, fraud detection rate in terms of false positive FP and true positive TP, methodology used and performance in terms of time required to produce the results (Bhattacharyya et al. (2011)). One of the most significant findings to emerge from the studies is that till now only sample of the large data sets were used to detect the fraud, so there is chance of false prediction, when there was no fraud and not predicting fraud, when actually fraud happened. Now in the Era of Big Data and new advance tools and technologies we can use the whole population data instead of only testing on small set to predict better results with more accuracy. But there is very less research found in the study of fraud using large datasets. This motivates us to extend this research work and develop a prototype of real time fraud detection application using advance data stream processing frameworks.

Bibliography

Agrawal, A., Kumar, S. & Mishra, A. K. (2015), 'Credit Card Fraud Detection : A Case Study Email Id :', pp. 31–33.

Bahnsen, A. C., Stojanovic, A., Aouada, D. & Ottersten, B. (2013), 'Cost sensitive credit card fraud detection using bayes minimum risk', *Proceedings - 2013 12th International Conference on Machine Learning and Applications, ICMLA 2013* **1**, 333–338.

Bhattacharyya, S., Jha, S., Tharakunnel, K. & Westland, J. C. (2011), 'Data mining for credit card fraud: A comparative study', *Decision Support Systems* **50**(3), 602–613.
URL: *http://dx.doi.org/10.1016/j.dss.2010.08.008*

Carminati, M., Caron, R., Maggi, F., Epifani, I., Zanero, S., Milano, P., Elettronica, D. & Bioingegneria, I. (2015), 'ScienceDirect B ANK S EALER : A decision support system for online banking fraud analysis and investigation', **3**.

Carneiro, E. M., Dias, L. A. V., Cunha, A. M. D. & Mialaret, L. F. S. (2015), 'Cluster Analysis and Artificial Neural Networks: A Case Study in Credit Card Fraud Detection', *2015 12th International Conference on Information Technology - New Generations* pp. 122–126.
URL: *http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7113459*

Chauhan, P. & Shukla, M. (2015), A review on outlier detection techniques on data stream by using different approaches of k-means algorithm, *in* 'Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in', IEEE, pp. 580–585.

Halvaiee, N.S & Akbari, M. (2014), 'A novel model for credit card fraud detection using Artificial Immune Systems', *Applied Soft Computing* **24**, 40–49.

Jiang, F. & Luo, D. (2014), 'A New Coupled Metric Learning for Real-time Anomalies Detection with High-Frequency Field Programmable Gate Arrays', *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on* pp. 1254–1261.

Kirkos, E., Spathis, C. & Manolopoulos, Y. (2007), 'Data Mining techniques for the detection of fraudulent financial statements', *Expert Systems with Applications* **32**(4), 995–1003.

Kumar, N. (2015), 'New Age Fraud Analytics: Machine Learning on Hadoop kernel description', `https://www.mapr.com/blog/new-age-fraud-analytics-machine-learning-hadoop`. Accessed: 2015-03-13.

Liu, C., Chan, Y., Alam Kazmi, S. H. & Fu, H. (2015), 'Financial Fraud Detection Model: Based on

Random Forest', *International Journal of Economics and Finance* **7**(7), 178–189.
URL: *http://www.ccsenet.org/journal/index.php/ijef/article/view/46957*

Mahmoudi, N. & Duman, E. (2015), 'Expert Systems with Applications Detecting credit card fraud by Modified Fisher Discriminant Analysis', **42**, 2510–2516.

Moro, S., Cortez, P. & Rita, P. (2015), 'Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent dirichlet allocation', *Expert Systems with Applications* **42**(3), 1314–1324.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y. & Sun, X. (2011), 'The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature', *Decision Support Systems* **50**(3), 559–569.

Phua, C., Lee, V., Smith, K. & Gayler, R. (2010), 'A Comprehensive Survey of Data Mining-based Fraud Detection Research', p. 14.
URL: *http://arxiv.org/abs/1009.6119*

Raj, S. B. E. & Portia, a. A. (2011), 'Analysis on credit card fraud detection methods', *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)* pp. 152–156.

Ravisankar, P., Ravi, V., Raghava Rao, G. & Bose, I. (2011), 'Detection of financial statement fraud and feature selection using data mining techniques', *Decision Support Systems* **50**(2), 491–500.
URL: *http://dx.doi.org/10.1016/j.dss.2010.11.006*

Rawte, V. & Anuradha, G. (2015), Fraud detection in health insurance using data mining techniques, *in* 'Communication, Information & Computing Technology (ICCICT), 2015 International Conference on', IEEE, pp. 1–5.

Sarno, R., Dewandono, R. D., Ahmad, T. & Naufal, M. F. (2015), 'Hybrid Association Rule Learning and Process Mining for Fraud Detection', (April).

Seemakurthi, P., Zhang, S. & Qi, Y. (2015), 'Detection of Fraudulent Financial Reports with Machine Learning Techniques', **00**(c), 358–361.

Sharma, A. & Kumar Panigrahi, P. (2012), 'A Review of Financial Accounting Fraud Detection based on Data Mining Techniques', *International Journal of Computer Applications* **39**(1), 37–47.

SIGKDD, A. (2014), 'DATA MINING kdd', `http://www.kdd.org/curriculum/index.html`. Accessed: 2014-03-13.

Zhou, W. & Kapoor, G. (2011), 'Detecting evolutionary financial statement fraud', *Decision Support Systems* **50**(3), 570–575.