FRAUD DETECTION IN MOBILE MONEY TRANSACTION: A DATA MINING APPROACH

By

Md. Shaheduzzaman Shahed

Id: 152-15-523,

Khalid Ibrahim

Id: 152-15-524,

Parvin Akter

Id: 152-15-576

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. S. M. Aminul Haque

Associate Professor

Department of Computer Science and Engineering

Daffodil International University

Co-Supervised By

Mr. Md. Reduanul Haque

Lecturer

Department of Computer Science and Engineering

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2019

APPROVAL

This Thesis named "Fraud Detection in Mobile Money Transaction: A Data Mining Approach", submitted by Md. Shaheduzzaman Shahed, ID No: 152-15-523, Khalid Ibrahim, ID No: 152-15-524, Parvin Akter, ID No: 152-15-576 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 5th May, 2019.

BOARD OF EXAMINERS

_____

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. S. M. Aminul Haque — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Saif Mahmud Parvez — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of Dr**. S. M. Aminul Haque, Associate Professor, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Dr. S. M. Aminul Haque
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Co-supervised by

_____

Mr. Md. Reduanul Haque
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Submitted by:

Md. Shaheduzzaman Shahed

ID: 152-15-523

Department of CSE

Daffodil International University

Khalid Ibrahim

ID: 152-15-524

Department of CSE

Daffodil International University

Parvin Akter

ID: 152-15-576

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

As a matter of first importance, we are so thankful and grateful to Almighty ALLAH for his kindness to enabling us to fulfil this research effectively.

Second, we would like to express our sincere thankfulness to our supervisor **Dr. S. M. Aminul Haque, Associate Professor,** Department of Computer Science and Engineering, Daffodil International University. We appreciate all his support and motivation through our research. His constant support inspired us a lot to complete this research.

We want to thank our parents who always trusted us, constantly supported and loved us, because without their love and support we would not be able to reach this level to complete our research.

We might want to thank every one of our classmates of Daffodil International University who also helped us and discussed with us while we were doing our thesis.

We also might want to thank the authority of Daffodil International University for giving us such a good environment and various facilities such as computer labs and library for using computer and discussing during our research.

ABSTRACT

In terms of all kinds of money transaction there occurs a common problem which is fraud. Fraud is an increasing concept that can badly affect to the economy over the whole world. There are various types of fraud such as mobile payment fraud, credit card fraud, bank fraud, insurance fraud and other financial frauds. So, it is important to prevent fraud from financial transaction to save the economy system. And predicting fraud is one of the efficient ways for preventing fraud. The objective of our research is to detect fraud in mobile money transaction using data mining. Actually, we will analysis a synthetic transactional dataset using classification and will predict the probability of fraud in future transaction.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Mobile money transaction has become a very popular form of money transaction nowadays. People all over the world use mobile payment system to make various type of payments. People of rural areas are also now using mobile payment system for various purpose such as, from buying product, transfer money to another account, cash out, to make recharge of others SIM etc. But the matter of concern in these terms is fraud transaction. Fraud in those transactions can affect to the economic condition of the normal people and also the whole country. Which can also affect the economic condition of the whole world. So, to prevent fraud in transaction fraud detection is an easy way. And we are doing this research to detect fraud in mobile money transaction using data mining. As the real transactional data is unavailable due to the matter of privacy and confidentiality, so we are using a synthetic dataset for doing our research. We have divided our dataset into two parts such as training set and test set and applied simulation to determine the accuracy of the test set. If the accuracy of the test set is fair enough then we can predict the fraud in the next transactions which can be very useful to prevent fraud in mobile money transaction.

1.1 Motivation

Frauds in money transaction can be a great threat for the economic condition of a country. So, to prevent fraud in money transaction fraud detection is a very effective way. But the matter of regret is that there is a few number research on this topic of fraud detection. And one of the most obvious reason is that as a matter of privacy and confidentiality there is also lack of real transactional data. So, our aim is to detect fraud in mobile money transaction using data mining. And we hope that this analysis will be very helpful to detect fraud and to prevent fraud from mobile money transaction.

1.2 Rationale of the Study

Mobile money transaction is one of the most common form of money transaction. But fraud in this transaction is a growing concern which can harm the economic condition of our country. So, fraud detection can be a productive way to prevent fraud in money transaction. A large number of researches on fraud detection can help to accomplish this goal. But there is a lack of real transactional data because of privacy of the user and as a result of that the amount of research is also very poor. So, our objective is to detect fraud in mobile money transaction using data mining. We also did not find any real data, so have used a synthetic dataset to perform our analysis. After dividing the dataset into two parts we looked forward to find the accuracy of the test set and based on that accuracy we can predict the possibility of fraud in new data and that can be very effective to prevent fraud. And we hope that the prevention of fraud will protect our economic condition from being decreased.

1.3 Research Questions

1. What will be the accuracy of the test data set?
2. How to detect fraudulent behavior in transaction?
3. How can we predict a fraud in money transaction?

1.4 Expected Output

We are trying to detect fraud from previous dataset and for this we have divided our dataset into two parts such as training set and test set. If the accuracy of the test dataset is impressive then we can apply the model to the new data and can predict the possibility of fraud in that new transactional data. So ultimately our expected output is the accuracy of the test set using Decision Tree, SVM and ANN and to visualize the decision tree.

1.5 Report layout Chapter

1. Background.
2. Research methodology
3. Experimental results and discussion.
4. Conclusion and implication for future research.
5. Reference.

CHAPTER 2
BACKGROUND STUDY

2.1 Related Works

1. Applying simulation to the problem of detecting financial fraud.
2. Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study.
3. A Comprehensive Survey of Data Mining-based Fraud Detection Research.
4. Learned lessons in credit card fraud detection from a practitioner perspective.
5. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature.
6. R. Riekeetal. "Fraud Detection in Mobile Payments Utilizing Process Behavior Analysis".In:2013InternationalConferenceonAvailability, Reliability and Security. IEEE, 2013, pp. 662–669.
7. D. Malekian and M. R. Hashemi. "An adaptive profile based fraud detection framework for handling concept drift". In: 2013 10th International ISC Conference on Information Security and Cryptology (ISCISC). IEEE, 2013, pp. 1–6.
8. C. Gaberetal. "Synthetic logs generator for fraud detection in mobile transfer services".In:2013 International Conference on Collaboration Technologies and Systems (CTS) (May 2013), pp. 174–179.

2.2 Research Summary

The summaries of research papers which we have read during our research are given below:

1. **The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature.**

Objective of the paper: This paper represents some technique of data mining algorithm which is discussed about financial fraud detection. For financial fraud detection 49 journal articles which published from 1997 to 2008 are analyzed. They identify four types of financial

fraud and applied six classes of data mining technique. For research they followed methodological framework, classification framework for application and analyzed financial fraud detection.

Algorithm/Method used by the article: Classification, Clustering, Regression, Visualization, Prediction, outlier detection.

Result: In this paper they worked on four categories of fraud (bank fraud, insurance fraud, securities and commodities fraud, and other related financial fraud). And they get different types of problems. And find out some limitations which creates several types of problem.

Future work: In this paper has two main limitations. Firstly, we used several keywords which published between 1997 and 2008. A future review could be expanded in scope. Secondly we write it in English and in future try to convert it different language.

2. A review of risk in banks and its role in the financial crisis.

Objective of the paper: The objective of this paper is to analysis the role of operational risk in the 2007/2008 financial crisis and to provide recommendations regarding the improvement of operational risk management to assist in the prevention of future crises.

Source of dataset used by the article: The dataset used by the article Esterhuysen at al. (2010).

Algorithm/Method used by the article: credibility theory, Value at Risk (VaR), Peak over threshold (POT), Hill's method.

Result: This research describes the 2007-8 financial crisis and Role of operational risk in the financial crisis and how should we act in any financial task. It also tells how we can improve operational risk management.

3. Learned lessons in credit card fraud detection from a practitioner perspective.

Objective of the paper: Due to fraud in credit card transaction there is caused a loss of billion dollars every year. So to reduce the losses, designing efficient algorithms for fraud detection can be an effective way. But the designing of these algorithms is very difficult for some reasons such as, dynamic distribution of data, incompatible distribution of classes and dynamic flows of transactions.

And there is also a lacking of real data for confidentiality and privacy matters. As a result we cannot be able to identify which is the most effective algorithm to handle them. So the objective of this paper is to generate some answers from the point of view of a practitioner by considering three critical issues: incompatibility, non-stationarity and assessment.

Source of dataset used by the article: The dataset used in this article is a genuine charge card dataset which they have got from their modern accomplice and that is an installment specialist co-op situated in Belgium. This dataset holds the logs of a subset of exchanges from the first of February 2012 to the twentieth of May 2013.

Algorithm/Method used by the article: Neural systems [12], Rule-based strategies (BAYES [13], RIPPER [14]), Tree-based calculations (C4.5 [15] and CART [16]), RF, SVM, NNET, inspecting technique (Under, SMOTE, Easy Ensemble, Incremental methodology (Static, Update, Forget).

Result: The paper exhibits the fraud discovery issue and proposes AP, AUC and Precision Rank as right execution measures for an extortion recognition task. The last best system executed the overlooking methodology together with Easy Ensemble and day by day update.

Future work: The programmed determination of the best unequal strategy on account of web based learning.

4. **Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study.**

Objective of the paper: Maximum time credit card is an easy way for fraud which takes short time and less risk. In this paper, supervise machine learning algorithm is used on real world datasets for detecting credit cards fraudulent transaction. Credit card datasets are very imbalanced dataset because it carries more allowed fraudulent transactions. The main goal of this paper is find out the accuracy and check the performance for the supervised machine learning algorithm.

Tools/Platform they used to analysis: In this paper, they applied ten machine learning models and compare their Accuracy, TPR, FPR, G-mean, Recall, Precision, Specificity and F1-Score. All machine learning algorithm is used for identity fraud or non-fraud transaction. The main purpose of this paper is to apply supervised machine learning algorithm to real world data sets.

Algorithm/Method used by the article: Supervised and unsupervised machine learning, unbalanced data, Fraud Detection Classifier.

Result: Described all models are giving better result in overall performance. So, top ten features can be used to find-out the accuracy, Recall, Precision, Confusion matrix and compare it to the old result.

Future work: In future they apply voting classifier and compare performance with other machine learning algorithms. They are also thinking to increase the training and testing dataset. Later on they use all the learning algorithms of machine learning and try to find out one of the best outcome.

5. Applying simulation to the problem of detecting financial fraud.

Objective of the paper: This thesis is for applying a monetary reproduction of two economic estate, mobile payments and retail stores systems. Because in every transaction there is a big problem called fraud, which can cause a failure in the economy. But because of the lacking of transaction data there is a poor amount of experimentation in fraud detection. So, the ultimate objective of this research is to apply a simulation in the detection of fraud and its application in the economic services. But as there is a lacking of real data, so they developed two simulators like mobile payment simulator (PaySim) and retail store simulator (RetSim) for generating synthetic transactional data and this data present both normal customer behavior and fraudulent behavior. They are also working on another simulator called *Banksim* which can be used for detecting money laundering cases.

The principle objective of building up these test systems is that it creates and share sensible and various fraud information with the exploration network.

Existing similar works & their objectives:

1. The work by Gaber et al. [21] presents another comparative procedure to produce manufactured logs for misrepresentation recognition.
2. Episode Response Sim by Gorton [22] is a reenactment instrument to help the appraisal of danger of web based financial administrations.
3. The work by Rieke et al., Zhdanova et al. [55, 66] on fraud recognition in portable installments is done in a comparable space as our work and different creators audited [21, 33].
4. Zhdanova et al. [66] is a continuation of the work done by Rieke et al. [55] and utilizes the test system created by Gaberetal. [21] To assess the outcomes.
5. Malekian and Hashemi [46] dealt with a fraud detection technique that handles the idea float on e-installments
6. Alexandre and Balsa [5, 6] present a technique to identify extortion utilizing wise specialists that play out the errands that physically a security officer ought to do without anyone else over a restricted measure of information.

Source of dataset used by the article: The dataset they have used for their research is a real transactional dataset which they have got from their exploration partner. This dataset was adjusted to coordinate the conduct of staff and clients utilizing accumulated exchanges from a store of one of the greatest shoe retailers in Scandinavia (Paper I).

Tools/Platform they used to analysis:

1. Retail Store Simulator (RetSim).
2. Mobile money Payment Simulator (PaySim).
3. Multi-Agent Based Simulation toolkit, called MASON.

Algorithm/Method used by the article: Verification and Validation.

Result:

1. Quantification and measurement of the quantity of loses committed by their noxious operators, this is particularly useful for estimating the expense.
2. Effectiveness of threshold detection.

Future work:

1. With the help of real data we want to improve the accuracy of the payment simulator PaySim.
2. Identifying complex kinds of frauds, for example, illegal tax avoidance
3. Modeling and improving BankSim by accessing real data sets.
4. Developing a multi-simulator by coordinating all three simulators that shares a typical reference to clients and can monitor the exchanges of a solitary operator over all test systems.

6. A Comprehensive Survey of Data Mining-based Fraud Detection Research.

Objective of the paper: The main objective of this paper is to identify challenges in different types of large data sets and streams. Then categories, compares and summaries relevant data mining-based fraud detection methods. This survey paper has been sampled by the last 10 years review paper articles. And also compare all related reviews on fraud detection which helps to take proper decision about FFD.

Source of dataset used by the article: Though this is a survey paper so they discussed about the different dataset used on the papers and analyses their attributes. They select four types of fraud (telecommunications, credit card, and insurance, internal) and made two.

Algorithm/Method used by the article: There are different types of algorithm mentioned in this paper which is used in many papers. They discussed which technique or method is given better result. They mentioned about four approaches,

1. Supervised Approaches on Labelled Data:
    - The neural network and Bayesian network
    - Decision trees, rule induction, and case-based reasoning have also been used
    - The cross validated decision tree
    - Two-stage rules-based fraud detection system
    - Case-based reasoning
    - Statistical modelling such as regression
2. Hybrid Approaches with Labelled Data
    - Supervised hybrid
    - Supervised/ Unsupervised hybrid
3. Semi-supervised Approaches with Only Legal (Non-fraud)
    - Kim et al (2003) applied in five steps fraud detection method on a novel
4. Unsupervised Approaches with Unlabeled Data
    - Applied unsupervised neural network method
    - Use cluster analysis for outlier detection, spike detection, and other forms of scoring

- Peer group analysis for inter account behavior
- Point analysis for inter account behavior over time
- experimental real-time fraud detection system based on a Hidden Markov Model (HMM

Existing similar works & their objectives: In this paper they applied different type of algorithm to find out fraud detection fraud such as credit card and telecommunications, and related domains such as money laundering and intrusion detection. Then outline techniques from credit card, telecommunications, and intrusion detection. Next neural networks, recurrent neural networks and artificial immune systems for fraud detection.

Result: This survey paper has covered almost all related studies about fraud detection. All types of fraud, methods and techniques are discussed here. After discovering the limitations about methods and techniques of fraud detection, this paper shows us that this field can benefit from other related fields.

Future work: In future work we want to work on the credit application fraud detection.

7. **A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers.**

Objective of the paper: Credit scoring and behavioral scoring are the two procedures by using which associations take decisions about to allow or to not allow the credit to shoppers who appeal to them in the hope of getting credit The objective of this review is to give an outline of the targets, systems and difficulties of credit scoring as an application of estimating. It also defines the method of changing the systems from determining the possibility of a buyer defaulting to deciding the profit a buyer will prompt the loaning association. It additionally brings up how effective has been this under-inquired about zone of determining financial hazard.

.

Tools/Platform they used to analysis: Demo-graphically based segmentation tool, graphical network tools.

Algorithm/Method used by the article: Algorithm/Methods used in Credit scoring:

1. Linear regression.
2. Recursive partitioning algorithm.
3. Logistic regression and classification trees.
4. Neural systems.
5. Expert frameworks.
6. Genetic calculation
7. Nearest-neighbor techniques.

Algorithm/Methods used in Behavioral scoring:

1. Bayesian Method.
2. Markov chains.

Algorithm/Methods used in Profit scoring:

1. Proportional hazards models.
2. Accelerated life models.

Result: Credit and behavioral scoring are the absolute most significant divining systems utilized in the retail and buyer finance territories As an unadulterated diving instrument instead of a basic leadership one, credit scoring has principally been utilized as a method for anticipating future awful obligation so as to set aside suitable provisioning. With the associations being made between scoring for default and scoring for focusing on potential deals, these scoring procedures will plainly be utilized to figure the offers of items just as the profit an organization will make later on.

2.3 Scope of the Study

Most of the papers we have read through our research have given only the review of some other papers. They didn't have done experimental work and didn't apply any algorithm to reach any decision. But in our research, we have directly applied algorithm to detect fraud in mobile money transaction. We have determined the accuracy of test set to make decision further on new data and visualize a decision tree that will help a lot to detect fraud and to prevent fraud.

2.4 Challenges

We tried to find real transactional data for our research of fraud detection, but we did not find any real data because as a matter of confidentiality transactional data are not available. So we started working with synthetic dataset. But the dataset is so huge that we have faced many problems to work with that dataset. At first, we tried to work with Weka (Waikato Environment for Knowledge Analysis is a suite of machine learning software written in Java). But as the dataset is huge Weka could not load the whole dataset. Then we started working with python. When we tried to visualize our decision tree, the tree was so huge that it was not possible to display clearly. After that, we tried to discretize all the attributes whose values are numeric but faced problems too because of the huge size of the dataset.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Fraud detection in mobile money transaction is a well-known problem. It's growing up day by day. Mobile money transaction is very much needed in developing countries where banking systems are not so much available. In many review papers researchers talked about different types of fraud and applied different algorithms in their individual research purpose. For our research we collected synthetic datasets from the source of Kaggle. There are eleven types of attributes. We use some data as training set and some for testing set. We used python software. Then applied transformation and reduction methods for the sake of our work. We applied classification for decision tree algorithm.

3.2 Research Subject and Instrument

The title of our research is "Fraud detection in mobile money transaction: A data mining approach". In this paper we use python language to implement our algorithm and use jupyter notebook as a platform.

3.3 Data Collection Procedure

Source of dataset: The source of dataset from Kaggle. Kaggle is a place that based on machine learning. Here's a discussion of data mining, datasets, data science along with machine learning. Many times, we get the training dataset and testing data set from the Kaggle to show us the kernel. Our dataset is a synthetic dataset that used for financial fraud detection. In the datasets the total fraud transaction is 8213. And not-fraud transaction is 6354407. The size of datasets is 471 MB.

Describing different attributes: In this dataset there are eleven types of attributes. There are given bellow:

Step: Step refers to the total number of data that can be passed through one medium in a single hour.

Type: There are five different types of transaction. They are PAYMENT, TRANSFER, CASH_OUT, DEBIT and CASH_IN.

Amount: Amount column presents the amount that the customer is going to transact in local currency.

nameOrig: This attribute represents the client who began the exchange/ transaction.

oldbalanceOrg: It represents the opening balance before the exchange/ transaction.

newbalaneOrg: This represents the new balance of the customer after the transaction.

nameDest: nameDest means the client who is the receiver of the exchange /transaction.

oldbalanceDest: This attribute describes, before the exchange /transaction how much balance does the recipient has. Here maximum initial balance is 0.

newbalanceDest: It represents the new balance after the transaction of the receiver.

isFraud: This attribute identifies whether the transaction is fraudulent or not. It contains 0 and 1. 01 represent fraud and 0 represent not fraud.

isFlaggedFraud: This attribute identifies illegal attempts to transfer more than 200000 in a single exchange / transaction. In our research we are ignoring this type of fraud.

Size of the dataset: This dataset contains huge amount of data. Total 6362620 data are present here. And total number of fraud 8312 and total number of not fraud 6354407.

Pre-processing: Data preprocessing is used to simplify the process of data processing. The main target of data processing is to find out the target or knowledge.

It is seen that there is a huge amount of data, there are some things we do not need, so data preprocessing is required. Besides, there are many problems in the data such as the data is inconsistent, incomplete and noisy.

Noisy means that data that you want to process in not accurate, bears miss information and is not complete. Data incomplete means I am adding a feature to forty students in my class, from where

40 students will have the first name last name. Many have seen the first name but did not give the last name and many have given the last name but did not give the first name. Due to this the data feature is not full.so data is incomplete.

Due to these reasons data preprocessing is required. Data preprocessing is our data mining fastener. And the algorithm works well.

There are four steps of data preprocessing. These are cleaning, integration, transformation and reduction. We have used only two pre-processing steps for our work convenience. Those are transformation and reduction.

Transformation: Transformation means to transform data from one format to another format. Normalization is another part of transformation. There are three types of normalization.

There are min-max normalization, Z-score normalization and decimal scaling normalization.

But in our dataset, we have to convert categorical data into numeric data using standard spreadsheet model.

For our datasets we used transformation for an attribute. 'Type' attribute is changed from categorical to numeric form. Though the dataset is synthetic, so the data is clean and there is no missing value as well. That's why we didn't have to do more preprocessing.

Reduction: Data reduction is transformation technique which create ordered or simplified form of meaningful data that derive from multitudinous amount of data.

For the sake of our work we have dropped some columns. These are step, nameOrig and is Flagged Fraud. For the benefit of our work, we have excluded these columns.

3.4 Methodology

Source of this dataset are collected from Kaggle. Our dataset is a synthetic dataset that used for financial fraud detection. The total number of data is 6362620. It's a huge amount of data. In the datasets the total fraud transaction is 8213. And not-fraud transaction is 6354407. The ratio of fraud transaction and not fraud transaction is 1:773. In every 773 transaction there is 1 fraud transaction happening. The size of datasets is 471 MB.

In this dataset there are eleven types of attributes. Every attributes means the different things. The attributes are step, type, amount, nameOrig, oldbalanceOrg, newbalanceOrg, nameDest, oldbalanceDest, newbalanceDest, isFraud, isFlaggedFraud.

Pre-processing: Data preprocessing is used to simplify the process of data processing. The main target of data processing is to find out the target or knowledge.

We use synthetic type of data.so no missing value or garbage value. Since there is no missing value here, there is no need for data cleaning. For these datasets we use only transformation and reduction method.

Transformation refers to transfer data one format to another format. For type attributes we transform categorical to numerical form. There is no change in other attributes.

Data reduction is transformation technique which create ordered or simplified form of meaningful data that derive from multitudinous amount of data. We have omitted some attributes for the benefit of our work. Step, nameOrig and isFlaggedFraud this three attributes are omitted for the work bases.

Classification is a procedure which is utilized to sort out and order information in various classes. It is a supervised algorithm because we are known about the class labels and the quantity of classes. Classification is a two-step process and the steps are model construction and model usage. In model construction step, each record of table belongs to a class which is determined by one of the attributes. These attributes are called target attributes and the values of the target attributes are called class labels. A model is learnt by using the set of all records and the model is called training set. In model usage step, we use a test set in which we skip the class attribute and produce the results ourselves using the knowledge we have learnt from the training set. Then we compare the result of test set with the result that we have got from the training set. If the percentage of accuracy is better than the previous result, then we can use the model for classifying new data. There are several classification methods, which are Decision Tree Induction, Bayesian Classification, K-Nearest Neighbor, Neural Networks, Support Vector Machines, Association-Based Classification, Genetic Algorithm etc.

For the sake of our research we used 70% data for training set and other 30% for test set and finally we get accuracy 99.97%.

Among all these methods we have applied three methods such as, Decision Tree, Support Vector Machines and Artificial Neural Networks and we have shown the comparative results of these three methods.

Decision Tree: A decision tree is like a flow chart. It represents tree structure. In a decision tree we test something and that test may have more the one result. This test is usually done on the attributes. It contains a root hub, branches and leaf hubs. Inward hub (non - leaf hub) presents test on characteristics, branch represents the out- come of the last, leaf node holds a class label.

When we have multiple candidate first split at that time there are multiple methods that one could use. The two well-known multiple methods are Information gain and Entropy. Entropy means disorder in a system. In a particular node all values are positive or all values are negative that is represent all example are the same class at that time entropy is 0 or entropy is low. On the other hand if the half value are positive and the half value is negative at that time entropy is highest.

When we choose the most useful attribute at that time one of the most useful criteria is information gain. Gain is measure how we reduce uncertainty (values lies between 0 and 1).

Let the arrangement of examples S (preparing information) contains p components of class P and n components of class N

The measure of data, expected to choose if a discretionary model in S has a place with P or N is characterized as far as entropy, I (p, n)

I (p, n) = - Pr (P) log_2 Pr (P) – Pr (N) log_2 Pr(N)

➔ Note that Pr (P) = p / (p + n) and Pr (N) = n / (p + n)

Information gain: If S_i contains pi cases of P and ni cases of N, the entropy, or the expected information needed to classify objects in all subtrees S_i is

$$E(A) = \sum_{i=1}^{v} Pr(S_i)\, I(p_i, n_i) \text{ where } Pr(S_i) = \frac{S_i}{S_i} = \frac{p_i + n_i}{p + n}$$

$$Gain\ (A) = I\ (p, n) - E\ (A)$$

for our research we are using decision tree methods to visit all the attributes of the data set to go to a specific decision.

Artificial Neural Networks: Artificial neural networks are computational models which is stimulated by biological neural networks, and used in generally unknown functions. ANN (Artificial Neural Networks) OR NN (Neural Networks) provide an exciting alternative method for solving a variety of problems in different fields of science and engineering. It is broadly connected in classification and clustering. The neural network itself is not an algorithm, but it's called a framework for many different machine learning algorithms to work together and process complex data inputs.it has some advantages

1) It is adaptive

2) It can create robust model

3) It can modify the classification process if new training weights are set.

4) Its real time operation

ANN are used in credit card, automobile insurance and corporate fraud.

Working process of artificial neural networks:

Start

Read the dataset

Encode the dependent variable (pre-processing o dataset)

Divide the dataset into two parts or training and testing

Tensor Flow data structure or holding features, labels etc.

Implement the model

Train the model

Reduce MSE (actual output-desired output)

Make prediction on the test data

End

Advantage:

->It's handle noisy and missing data

->It can work with large number of variables or parameters.

->It provide general solution with good accuracy

-> It deal with the non-linear function

-> System has got property of continuous learning

For our research we applied ANN algorithm to find out good accuracy. Firstly ANN algorithm train an amount of data then applied test rest of the data. After that we find a good accuracy.

SVM: SVM means Support Vector Machine. It is a supervised learning methods. SVM used associated learning algorithms that analyze data used for classification and regression and other or outlier detection. So SVM is a supervised machine learning methods that's looks at data and sorts it into one of the two categories. Support vector machine is one of the most effective classifiers which have sort of linear. It has a very you know good mathematical intuition behind the support vector machine and we are able to handle certain cases where there is non-linearity by using non-linear basis functions or in particular we will see these are called kernel functions.

Why support vector machine is so popular?

We will see that support vector machine have a clever way to prevent over fitting. And we can work with relatively larger number of features without requiring too much computation

Flow chart o SVM algorithm:

Prepare and format dataset

Normalize dataset

Select activating functions

Optimize parameters and using search algorithms after cross validation

Train SVM network

Test SVM network

Evaluate model performance

SVM is find out a separating line or hyper plane between data of two classes. It takes data as an input and outputs a line that divided those classes If possible. It is used or pattern recognition, speech recognition, face detection, faulty card detection.

Application:

- ➔ Bye the help of kernel unction we can solve any complex problems.
- ➔ The risk of over fitting is less in SVM
- ➔ Compare with ANN, SVM give better result.

Disadvantage:

- ➔ Kernel unction is no easy
- ➔ It takes long training time for large dataset

Application:

- ➔ Really good or text classification.
- ➔ Handwriting recognition
- ➔ Protein structure prediction

- ➔ Breast cancer diagnosis
- ➔ Intrusion detection

SVM is a machine learning algorithm. For our research at first we takes some data as a training data set then rest of the data are used for test data set. Then find an accuracy .But SVM takes long training time for large data set. But the accuracy is better than ANN.

Confusion Matrix: Confusion matrix represents a table which is applied on a test data set to analyze the performance of a classification model or classifier .There are two parts of confusion matrix one is predicted part and another is actual part. Actual value represent which value is true from previous stage and predicted value represents after experiment or observation we have to say something as like as value true or false.

Key matrix:

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

Recall = $\frac{TP}{TP+FN}$

Precision = $\frac{TP}{TP+FP}$

TP =True Positive

TN = True Negative

FP =False Positive

FN =False Negative

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Result

- Our dataset is synthetic dataset that use for money transaction fraud detection. The total number of data is 6362620.
- Data taken 10,020 from 6362620. Fraud = 1010
- Training set 7014 (70% of taken data). Fraud = 702
- Rest is test set (30%)

Result type	Decision tree	Support vector machine (SVM)	Artificial neural network(ANN)
Confusion matrix	[[2695 3] [176 132]]	[[2693 5] [111 197]]	[[2674 24] [41 267]]
Accuracy	94.04524284763806	96.14105123087158	97.83765801729874
Precision, Recall	precision recall 0 0.94 1.00 1 0.98 0.43	precision recall 0 0.96 1.00 1 0.98 0.64	precision recall 0 0.98 0.99 1 0.92 0.87
Executing time(sec)	263.8233003139994	0.49369112200020027	91.12833088299976

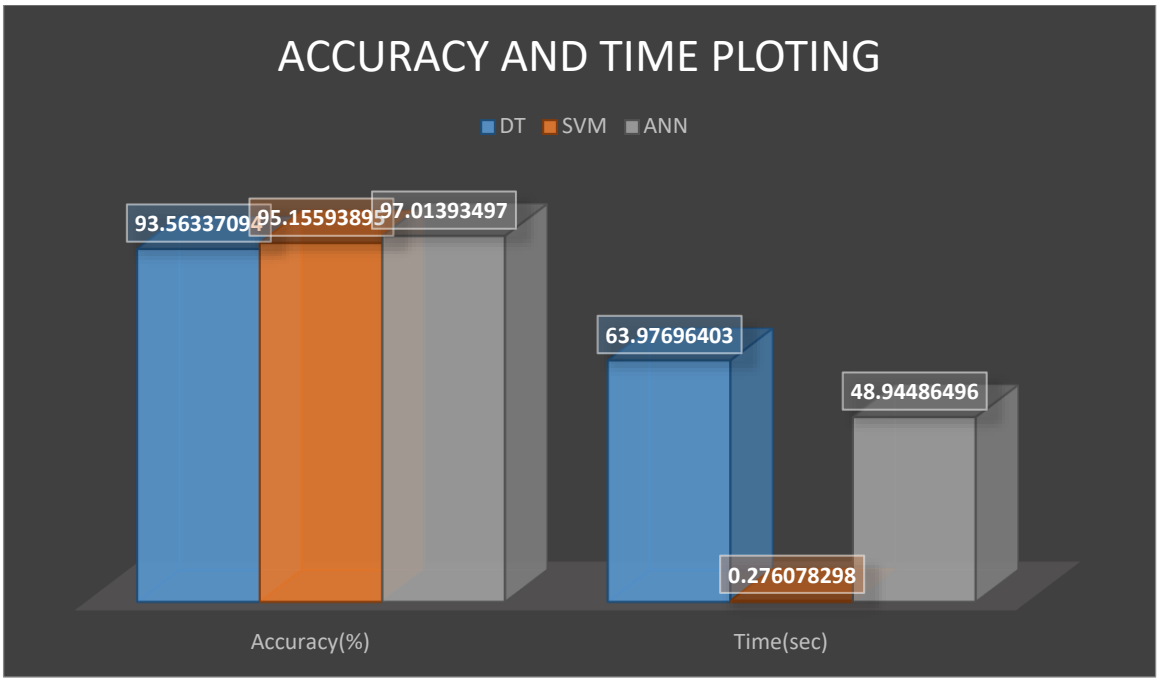

Figure 4.1.1 Accuracy and time plotting

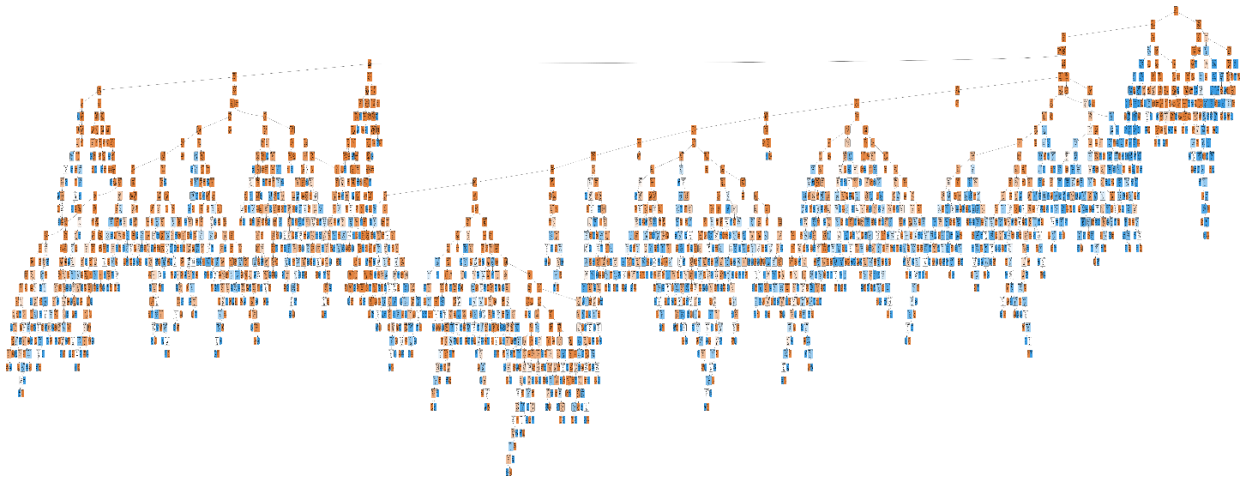

Figure 4.1.2 Decision tree of the selective dataset

4.2 Descriptive Analysis

For our research we applied three types of algorithms. This are decision tree, support vector machine algorithms and artificial neural network algorithms. Among those three algorithms we try to find out the best accuracy and executing time. Our total data is 6362620. This data type is synthetic. It's a huge amount of data. This large amount of data run very tough and it takes more time to executing. Firstly we use 5022 data for test from 6362620. Here the amount of fraud data is 574 and not fraud 4448. And also use 3515 data for training set (70% of taken data). Here fraud is 400 and not fraud data is 3115. And the rest of 30% data is used as a test set.

For decision tree accuracy is 93.56%, support vector machine accuracy is 95.15% and the artificial neural network accuracy is 97.01%. So here artificial neural network algorithms gives better accuracy then others two algorithms.

Secondly we use 10020 data for experiment .among 10020 data fraud is 1010 and not-fraud is 9010. We use 70% data for training set. So 7014 data is used for training set here fraud is 702 and not fraud is 6312. And rest 30% data is used for test set.

Decision tree: we used 10020 data for experiment. And after the experiment we find the accuracy is 94% and executing time is 263.8233 sec.

Support vector machine: when we applied the support vector machine algorithm the accuracy is 96%and executing time 0.493 sec.

Artificial neural network: After applying artificial neural network algorithm the accuracy is 97.83 and executing time 91.12 sec.

Among this three algorithms we see that artificial neural network algorithm gives better accuracy 97.83 but executing time is 91.12 sec. On the other hand support vector machine algorithm gives the accuracy is 96.14 and the executing time is 0.493 sec. SVM takes very short time for execution.

Among the two types of data quantity we find that artificial neural network gives better accuracy.

CHAPTER 5

CONCLUSION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Conclusions

Fraud in mobile money transaction we understand cash-in, cash-out, mobile research, national and international transfer, bill payments etc. Fraud in mobile transaction is increasing day by day. If we do not take action now, then it will have a huge impact. Because of these billions of dollars go to the hands of fraudsters. This research gives us the idea which type of transaction is fraud transaction in mobile money transaction

5.2 Implication for Future Study

For any kind of research in data mining the first challenge is collecting data. Ours is not different. We have to face many problems for collecting data. For example, privacy issues or organizations rules etc. So, we have to go for synthetic data. But in future we will do research on the real-world data. It will help to predict the real fraud and its behavior in the financial service.

CHAPTER 6

REFERENCES

[1] A Comprehensive Survey of Data Mining-based Fraud Detection Research [1] .

[2] Applying Simulation to the Problem of Detecting Financial Fraud [2] .

[3] A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers [3].

[4] Learned lessons in credit card fraud detection from a practitioner perspective [4].

[5] The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature [5] .

[6] Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study [6].

[7] https://www.kaggle.com/ntnu-testimon/paysim1/discussion/31087?fbclid=IwAR1QR_NkNptuUQDN8FfZSpygRNcepZ3WJvn3VAorDKrb-tkC5b-4mGKjDD4

[8] https://en.wikipedia.org/wiki/Kaggle

[9] file:///C:/Users/PARVIN/Desktop/Classification-Decision-Trees.pdf

[10] https://pythonprogramming.net/introduction-python3-pandas-data-analysis/

[11] https://github.com/bhattbhavesh91/visualize_decision_tree