Comparative Analysis of Data Mining Techniques for Fraud Detection (A Case Study of Branchless Banking)

M. Talha Umair
MS Computing
Shaheed Zulfiar Ali Bhutto Institute of Science and Technology
90 and 100 Clifton
Karachi -75600
talhaumair@gmail.com

Dr. Syed Saif-ur-Rahman
Assistant Professor, Faculty of Computing
Shaheed Zulfiar Ali Bhutto Institute of Science and Technology
90 and 100 Clifton
Karachi -75600
saif.rahman@szabist.edu.pk

Abstract— Data mining algorithms have been using since few years in financial institutions like banks, insurance organizations, etc, and these organizations are using applications of data mining techniques in prediction of business collapse, marketing analysis and fraud detection. In this study our objective is to provide a comparative analysis and find the most suitable techniques of data mining for fraud detection in the area of branchless banking on certain comparison criteria. We have used few different mining algorithms like decision tree, association rules, clustering, naïve bayes and neural network. Our other objective is to find out the comparison criteria, through which we compare these algorithms and that criteria are training volume (small dataset) against quality patterns level, model creation Time, ease of implementation, ease of presentation, extensibility, efficiency, simplicity, training volume (large dataset) against quality patterns level, popularity. In the end we have suggested the most suitable algorithms for fraud detection on branches bank.

Keywords—*Fraud, Data Mining, WEKA, SSAS, Decision Tree, Clustering, Association, Naive Bayes, Neural Network*

I. INTRODUCTION

A fraud is an intentional cheating, purposeful misrepresentation in order to get gain personally or to damage another people. In the past, many financial organizations have faced fraud activities and still they are facing because most of the time, fraudular uses different ways to exploit the rules and regulation to gain money or damage another person. Fraudular don't use same techniques, they use different technique to exploit the regulation or to do fraud. There are many different types of frauds, from false identity, insurance fraud, tax fraud, and making false statements etc. If we specifically focus on banking area, we have seen even from the past and still we sees that fraud happens in different areas of bank like credit card, net banking, branchless banking etc, [AS07]. In this research, our purpose is to provide a comparative analysis of different data mining techniques for fraud detection and to explore and suggest most suitable data mining techniques in the area of branchless banking.

In recent times, branchless banking has developed with very fast pace around the world. People use their services with those devices, which are available cheaply and also used by almost every one for example mobile phones. People as branchless bank's customers or agent can use their mobile to perform different transactions. There is no time restriction in branchless banking in terms of performing transactions. Being a new financial area, there is chances of flaw in terms of technical privacy and security. Sometimes Agents do unusual and suspicious activities when they performing transactions in prohibited way according to the regulation. Our objective is to compare the different data mining techniques and find the most suitable technique(s) in the area of branchless banking. Out suggested technique will definitely help to minimize and reduce security laps and fraud concerns in branchless banking. In this study, first we have tried to identify those parameters or attributes which can help to give more realistic patterns and also help to find out frauds cases from normal cases and then we applied different data mining techniques like neural network, naïve Bayes, decision tree etc., on our data and compared results of all these techniques on certain comparison criteria like training data volume, model creation time, model efficiency (fast or slow) etc., and after comparison, we have suggested the most suitable techniques for fraud detection in the area of branchless banking.

This study is an extension of our previous work that was 'Fraud detection using data mining on branchless banking' [MT12]. In that research, we have done fraud detection using classification technique that is decision tree in which first we collected data from branchless bank, labeled it, select appropriate attributes and applied decision tree. We have used a data mining tool named 'WEKA' and we used different algorithms like random tree, j48 etc. After doing some experiments, we have concluded that random tree is the most appropriate algorithm for fraud detection on branchless banking [MT12]. So in this current research we have further applied different data mining techniques like neural network, naïve Bayes etc., on our data and we did a comparative analysis using certain comparison criteria and in the end we have suggested more suitable techniques.

II. BACKGROUND KNOWLEDGE

This section contains basic idea about Branchless banking, data mining, different data mining techniques, which are necessary for the study.

A. Branchless Banking

Branchless banking is a financial firm which offers different financial services with no connection, association or participation of bank. Branchless bank offers different channel like Mobile, SMS-Command, Internet-Agent and many more. Most frequent channel on daily basis used by branchless banking is Internet-Agent. Millions of transactions are performed within a week by this channel. It also generates the most of the revenue among all channels. This channel is used by normal shop keepers. There are many transactions, provided by this channel particularly like paybill, eZLoad, pay to any one (other person must be customer or agent of that branchless bank), pay to CNIC, cashin, cashout, etc. Figure 1 show the architecture of branchless bank where branchless bank switch is attached with different channels and different cellular links.

Branchless bank connected with customers (using SMS and IVR channel) and with agents (using Internet, GPRS channel). Different telecommunication companies like ufone, zong also exposed their services to the branchless bank. 1-link is one of the important core components which connect the branchless banks to formal banks. For instant, if customer perform a transaction to buy air time of ufone (eZLoad), then bank first check the customer's account regarding different validation like balance enquiry, limit validation, etc., then check the telecommunication company 'ufone' is available. After all the validations system will deduct the amount from customer's account and then share the air time to his provided mobile number. If transaction has any kind of tax, then system will deduct the tax from customer's account. If this transaction is done by agent then certain commission will be credited to agent's account.

B. Data Mining

Data mining is extraction of knowledge, show hidden patterns; expose hidden answers which are previously unknown, from the data. In this section we have discussed some techniques and their concepts which are needed to know about our study.

C. Classification

One of the popular techniques of data mining is 'Classification', where data (input attributes) is organized, classified according to the target attribute (predicted attributes). Classification can be performed with the following techniques:

- Naïve Bayes
- Neural Network
- Regression
- Decision Tree

A. *Naïve Bayes*

A naïve Bayes classifier is also one the technique to construct classification and it presumes that the presence or absence of a particular element of a class is unrelated to the presence or absence of any other element for given class variables.

B. *Neural Network*

An Artificial Neural Network (ANN), usually known neural network (NN). It is a computational model derived from biological neural networks, consists of group of artificial neurons or nodes with an interconnected and they processes information using a connectionist approach to computation. It is generally used in complex problems to find patterns from data.

D. Clustering

Clustering is one of the technique in which data is classified, grouped according to the those groups, whose having almost same behaviour, same characteristics and attributes value. It is also known as 'Segmentation'. It is an unsupervised technique. The only different between classification and clustering is that clustering cannot use for prediction.

E. Association

Association is one of the most popular data mining techniques. Association generally uses to target sales problem like 'market basket problem', where one product is associated with another product. In other words association identifies common set among the two objectives, first is to identify the frequent objects sets and then identify association rules.

III. Related Work

Adrian et al. provided the comparative analysis of different data mining technique like decision tree, artificial neural network, logistic analysis, survival analysis etc., on the area of automotive insurance fraud detection [AG12]. They compared results of each technique and conclude that all the techniques are important in the area of insurance fraud detection because their objective was to cover the important issues like, it should not predict or detect legitimate customer as a fraudular and it should detect maximum fraud cases to avoid big loss for an organization [AG12]. At the end, they suggested decision tree as the effective technique in fraud detection and they also suggested that neural network for fraud detection but it requires big data. They suggested that when ever try to implement models to detect automotive insurance fraud, first consider issues in specific cases like resource constraints and another thing to take staff in a loop to keep in the process in order to get benefit from their better ability to handle the constant change in the field.

Kate smith et al. did a survey based research on fraud detection via data mining where they took all the papers from last 10 years related to the problem [KS10]. They defined the different types of fraud, sub types of fraud, the practical nature of data, performance metrics, methods and techniques. They analyzed and compared these techniques using some criteria that certain techniques are suitable for certain area like unsupervised approach is suitable for counterterrorism work, monitoring system and text mining from law enforcement and semi supervised from spam detection [KS10].

Afshar Alam et al. analyzed on different association rules algorithm [MA11]. They have performed experiments on four algorithms that were AIREP, Scaled rules, FP growth and Apriori [MA11]. They compared these entire algorithms on different comparing criteria like different type of datasets, support, number of rules produce, etc. Scaled and AIREP generated almost the same frequency of rules on each type of dataset like in smaller dataset, normal dataset or large dataset. They experimented on real world datasets. These two algorithms showed maximum number of rules but still there are chances to had lots of unwanted rules as well. So we took inspiration from their work and we considered their comparing criteria for our different data mining algorithms. They are using

different algorithms of association rules while we used different data mining algorithms like classification, clustering, association, etc.

In another research work, author has evaluated the different data mining software applications /systems and they proposed to compare them on the basis of quality attributes [EC12]. They compared data mining systems on the basis of quality attributes because every application have their own architecture, different way of use, ease of use etc. They have taken different data mining systems like High performance data mining system, The Quest, WekaG, UMiner, DBMiner, Anteater etc., and they took some non functional requirements or quality attributes as comparison criteria like Correctness, Extensible, Flexible, Integrity, Efficient, Privacy Preserving, Customizable, Usable, Transparent, Comprehensiveness, Multidimensional Dimensional Data, Large Data. In the end they have concluded and proposed that a parallel architecture for distributed data mining systems fulfilled all the requirements and demands of these systems. This would provide a comprehensive system which is speedy, portable, parallel, data /system transparent, protected, customizable, usable, extensible, and flexible that keeps large data [EC12].

IV. Research Methodology

The objective of this work is to compare different data mining techniques and explore the most suitable technique(s) for detecting the frauds. We have used different data mining techniques like association rules, clustering, naive bayes, neural network and decision tree for analysis and comparison. This section contains steps of data mining, like data collection, training data set, selection of appropriate attributes, different data mining algorithm. In the end we have some comparison criteria and comparison factors and on the basis of those criteria and factors, we are able to suggest the most suitable technique(s) for fraud detection on branchless banking.

F. Collection of Data

As we already mentioned that it is extension of our previous research study, so we are using the same data that we used in our previous study. We also tried to find out some other relevant data from customer's table and also from transactional table, but unfortunately, we did not able to extract further useful data from that branchless bank, which can help us more. In this study, we are also able to manage long data set, which is almost over 4500 rows and once again we got this data from the same branchless bank.

G. Size of Data

We have got small dataset from branchless banking which includes all the five transaction types which we have focused. It consists of 500 rows. Each row is labelled by 'normal' or 'fraud'. We also got long data set which consists of over 4500 labelled rows.

H. Structure of Data Set

We have got a data from transactional table and it is consist of 10 attributes. 'Tag' is a additional attribute which classified the data weather as normal or fraud. The attributes from the provided table and was collected from the branchless bank, are given below:

Attributes
Transaction id
Transaction type
Channel
Amount
From account number
To account number
Transaction date and time
Customer type
Commission type
Tag

I. Relevant Attributes Selection

We did not include all the attributes which we have got, so we have picked only those attributes which are relevant for our research study. For instant, we have not included 'FromAccountNo' and 'ToAccountNo' directly; we used them in summarized attributes like 'TotalTxnPerDay' between similar 'FromAccountNo' and 'ToAccountNo'. Reason behind for this attributes reduction is to make models more generalize rather than account numbers specific. Same logic also applied for 'TransactionDate' as well. We used 'TransactionDate' in summarized attributes. A part from these three attributes, we have used all remaining attributes in our study.

Another thing we have done, we have created two columns to make out data more comprehensive and following are those two added attributes:

- "TotalTxnPerDayF": Sum of all transactions for the same day, having same debit and credit account, same TxnType, same TxnChannel, same CustomerType
- "TotalTxnPerMonthFT": Sum of all transactions for the same month, having same debit and credit account, same TxnType, same TxnChannel, same CustomerType
- 'TxnType' stands for transaction type, TxnChannel for transaction type.

J. MS SQL Server 2008 Analysis Services

It is also known as 'SSAS' and used for several purposes like it can be used for data mining, BI (business intelligence) projects and data warehousing projects. SSAS offers different data mining techniques which are given below:

- Microsoft Decision Trees Algorithm
- Microsoft Association Algorithm
- Microsoft Naïve Bayes Algorithm
- Microsoft Clustering Algorithm
- Microsoft Neural Network Algorithm

We have already used Decision Tree in our previous research study using WEKA as a data mining tool, so we did not use Microsoft Decision Tree. In this research study, we have used Microsoft Association, Microsoft Naïve Bayes, Microsoft Neural Network and Microsoft Clustering for comparison. A reason for selecting MS SSAS 2008 is because, it has some advantages over other tools like it has a comprehensive GUI and of course it is a Microsoft's product and they always provide those applications which are easy to use, easy to understand and also we did not want to serve our time in tool exploration. So we just studied few tutorial for getting understanding, how to use this tool and started data

analysis. Data should be in the form of table in a DBMS (MS SQL).

K. Comparative Analysis

We have studies various research paper regarding comparison of different data mining techniques for fraud detection. On the basis of those papers we have finalized few comparison criteria or comparison factors to compare from different data mining techniques. Following are the list of comparing criteria's:

- Training Volume (Small dataset) against Patterns Level: Give better results in terms of good patterns, when there is small dataset provided.
- Model creation Time: How much algorithm takes time from start to end (provides relevant inputs and gets reasonable output).
- Ease of implementation: How much algorithm is easy to implement.
- Ease of Presentation: How much easy to understand and present results (patterns) to yourself and to others as well.
- Extensibility: The ability to algorithm to give results from small amount of data to large amounts of data sources.
- Efficiency: Give better results in noisy data
- Simplicity: it refers to the simple structure (output) of the algorithms and has adaptability by any person
- Training Volume (Large dataset) against Patterns Level: Give better results in terms of good patterns, when there is large dataset provided.
- Popularity: Which algorithm you have found or studied popular, in most of the problems solving situation.

L. Scoring Criteria

We have set the score criteria over comparison criteria on different data mining techniques which are following:

Score	Description
1	Poor
2	Fair
3	Good
4	Very good
5	Excellent

M. Users Survey

We have also conducted a survey according our research study that is 'comparative analysis of different data mining techniques'. Survey users are the students of BS and MS in computer science, software engineering and they have studied all these data mining techniques in a 'data mining' course. All students have basic knowledge and understanding of all those techniques which we have consider in this study and they can rank the data mining techniques accordingly. Basically this survey is based on experiences and expertise of the people/students, whether in professional work or academics tasks. Purpose of this work to know, the comparison of our final ranking of all the algorithms with the other's ranking of all the algorithms. It gave us inspiration of our work. Survey questions are the same as we mentioned in section 4.6 and scoring criteria also is same as section 4.7.

V. Discussion

In this research study, we have tried to identified some more attributes related to agent or customer which can really help us to give much more meaningful patterns, but unfortunately we could not able to find further more attributes. As we already have discussed earlier that, this is the extension of our previous work and we already put efforts to find maximum number of attributes for fraud detection using data mining. We have once again contacted to our branchless bank in order to explore and analysis for relevant attributes selection. We have explored on transactional table as well as customer personal information table, where we did not find any helping attribute. So in this study, we are using the same schema as we use in our previous study.

Another important thing is that, we have used WEKA for data mining, which gave us lots of algorithms for decision tree [WE12]. But in this study we have not used WEKA, we used MS SQL 2008 SSAS for data mining. We have tried different tools like WEKA, Orange, Rapid Miner and SSAS for different algorithms like classification, clustering, etc. After our rigorous effort, we came to conclude that SSAS provides better presentation of results, ease of use and ease of understanding. It is much more mature than other tools and there are lots of examples for getting understanding over the internet.

VI. Experimental Arrangements

We have used MS SQL Server 2008 Analysis Services (SSAS) for data mining and MS SQL Server 2008 for DMBS. SSAS provides some data mining algorithm, which was good enough for us to achieve our goals. In the experimental setup, first we imported table that contains transactional data of branchless banks having both normal and fraud cases into MS SQL Server, then we created some purposeful views to get countable (summarized) information from different attributes like 'sum of all transactions for the same day, having same debit and credit account, same TxnType, same TxnChannel, same CustomerType' and 'sum of all transactions for the same month, having same debit and credit account, same TxnType, same TxnChannel, same CustomerType'. After finalizing our view, we import it on SSAS as the data source.

Figure 1 shows the parameters, which is set of input, predict, predict only, key and ignore parameter. SSAS provides different data mining analysis on a single project.

Fig. 1 - Input parameters for data mining techniques

A. *Normalization*

We have transformed our data (all attributes) into normalized data, because in clustering, native bayes and neural network, we need normalized data for input attributes and predictable attributes. It is an important step of data pre-processing in knowledge extraction process. Normalization can be done through various methods like min-max or z-score. We have used min-max normalization. Formula for the min-max normalization is given below:

$$B = \left(\frac{(A - minimum\,value\,of\,A)}{(maximum\,value\,of\,A - minimum\,value\,of\,A)}\right)$$

Where 'A' is a current value, 'minimum value of A' is minimum value of the attribute and 'maximum value of A' is maximum value of the attribute. Output of normalization is always between '0.0' to '1.0'.

B. *Association Analysis*

When we applied Microsoft Association on our data, we got my combination of different association rules for both normal cases and fraud cases. Since we are focusing on fraud analysis, so our main concerns is with fraud cases only. When we applied some filter in the form of regular expression, and it was specific for fraud cases then it was just showing association rules for fraud cases only. In section appendix, figure 2 shows the filtered association rules for fraud cases. In this way, we have got many association rules, which can really help us to identify the suspicious transactions. Although it is showing few incomplete rules, so we ignored them and picked the most relevant association rules according to our objective.

C. *Clustering – Outlier Analysis*

When we have applied clustering on our normalized data, we got the fraud outliers from the normal clusters. Cluster 6, 7, 8 and 10 highlighted with dark blue shows that there is a maximum number of chances of fraud pattern in that all four outliers, while remaining are the normal clusters. In figure 3, we have marked all four fraud outliers with red circle to make it visible as fraud outliers, while remaining are the normal clusters. Our objective is not to focus on normal transactions, so we have just ignored normal clusters. While cluster 2 has little chances of having fraud cases, so we ignored it as well.

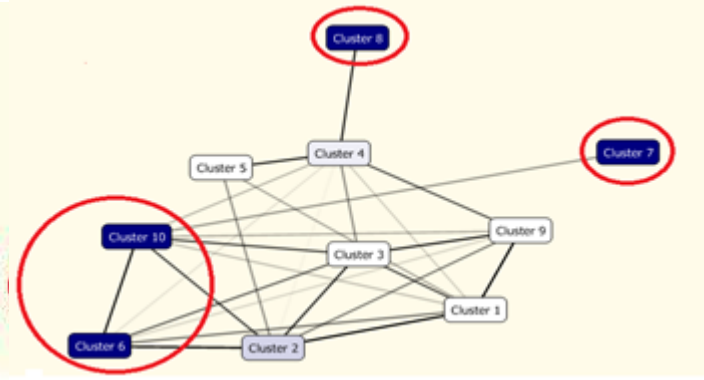


Fig. 3 - In clustering, fraud outliers identified from normal clusters that is cluster 6, 7, 8 & 10

If we go in details at clusters level then figure 4 shows the cluster 7 and it is stating that if transaction type is 'cashin', channel is internet, total transaction per month for the same to account, same from account and same channel are greater than 13, sum of all transactions per day for the same things are greater than 11 and commission type is fixed on each transaction, then there is a chances of fraud case. It means agent is doing 'cash deposit' on customers account in multiple slabs in order to make his commission multiple to the numbers of transactions instead of single transaction on 'cash deposit', which makes it suspicious

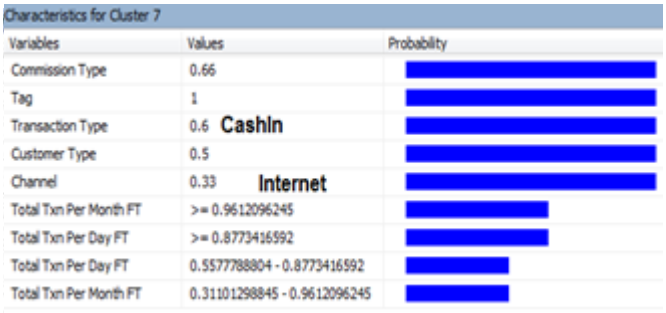


Fig. 4 - Cluster 7, a fraud outlier

D. *Naïve Bayes Analysis*

When we have applied naïve bayes on our normalized data, we did not get the same or at least closer dependency network compare to others algorithm, like we got in association rules. Figure 5 shows the dependency network of naïve bayes after applying our data on it. It should show all the attributes around the tag node. Although we have provided the same input and predictable attributes that we provided in others algorithm, just two nodes formed around the tag node that is 'total transaction per month for same to account, same from account and same channel' and 'total transaction per day for same to account, same from account and same channel'. But somehow this graph shows the relationship between these nodes, which may be critical as far as fraud case patterns identification.

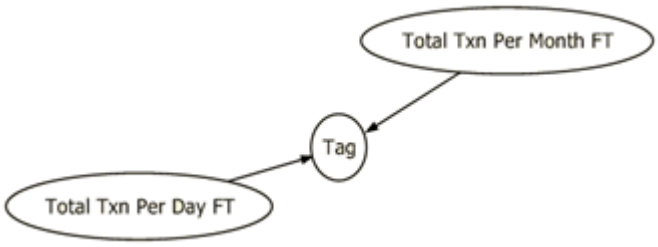


Fig. 5 – Dependency Network of Naïve Bayes

Figure 6 shows somehow patterns but it is not showing patterns specifically with respect to transaction type, customer type or commission type, etc. It is just showing that 'sum of all transactions for the same month, having same debit and credit account, same txntype and txnchannel' is greater than 11 then there is a changes of fraud cases or if 'sum of all transactions for the same day, having same debit and credit account, same txntype and txnchannel' is equal to 1 and total transaction for same things are between 11 and 18 then there are changes for fraud cases .So we consider it as an unacceptable algorithm for our branchless data and it is not acceptable for our research work.

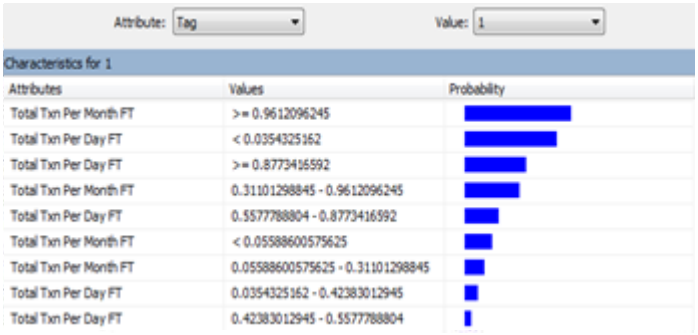


Fig. 6 - Attribute Characteristics of Naïve Bayes

E. Neural Network Analysis

When we applied our normalize data on neural network and apply filter specific to transaction type, we came to know that patterns are repeating in every transaction type. As we know that neural network give better result on large dataset and we have small data set, so it may be resulting to produce unwanted patterns from our data using neural network.

Figure 7 shows that output results of artificial neural network, specific to 'send' transaction type. Good part of this algorithm that it is showing the good trends and patterns of fraud cases against all input attributes but problem is clearly shown that when we selected other transaction type, same repeating patterns shown as well.

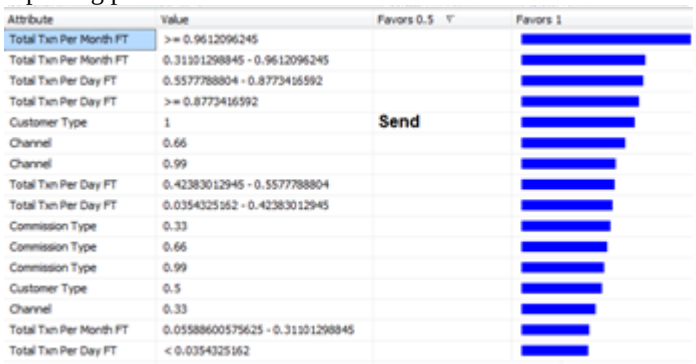


Fig. 7 - Neural Network analysis for Send transaction type

We can clearly and easily observe that similar trends and patterns are showing with similar input attributes in both figures 18 and 19. Even we have selected different transaction type like 'send' and 'load', almost similar output is showing.

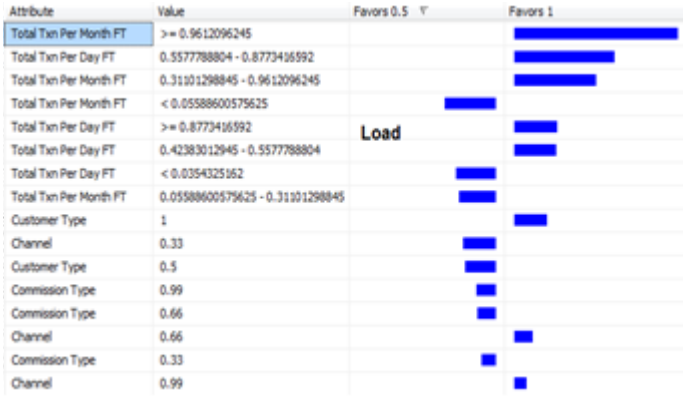


Fig. 8 - Neural Network analysis for Load transaction type

F. Decision Tree Analysis

In section appendix, figure 9, shows the decision tree, which we have achieved by applying random tree algorithm on our branchless bank's data using WEKA mining tool. We already mentioned that, this is our previous work of Independent Studies - 1.

In our previous study, we did not just perform experiment on our data using random tree algorithm, we have done experiments for almost all of the different algorithms of decision tree using WEKA data mining tool. We picked the best four algorithms after some analysis, observation and experiments on all algorithms. We did some experiments and on the basis of those experiments, we have concluded that random tree is a most suitable decision tree's algorithm for branchless banking.

In section appendix, figure 9 shows the braches of transaction type 'Send' and it was showing that if channel is internet, total transaction per month for same debit and credit, and same channel' is greater than 6.5 than there is a chances for fraud cases. Figure 9 also shows the branches of transaction type 'CashIn' and 'Load and it is stated that if total transaction per month for debit and credit account and same channel' is greater than 6 with transaction type 'CashIn' then there is a changes of fraud cases and if total transaction per day for debit and credit account and same channel' is less than 2.5 with transaction type 'Load' and total transaction per month for debit and credit account and same channel' is greater than 10.5 than there is a chances of fraud cases. Figure 9 shows the branches of transaction type 'Topup'. It is showing that if total transaction per day for debit and credit account and same channel' is between 1.5 and 2.5 with transaction type 'Topup', channel 'SMS' and total transaction per month for debit and credit account and same channel' is greater than 6 then there is a chances of fraud cases.

In the end we have concluded in our previous study that random tree is most suitable algorithm among all decision trees algorithm for fraud detection on branchless banking.

G. Comparative Analysis

Comparison Criteria	**Data Mining Techniques**				
	Decision Tree	Association	Clustering	Naïve Bayes	Neural Network
Training Volume (Low size) against Patterns Level	4	3	2	1	1
Model creation Time	3	4	2	2	2
Ease of implementation	4	4	2	2	2
Ease of Presentation	5	3	2	1	1
Extensibility	3	4	3	2	2

Efficiency	4	4	4	2	2
Simplicity	4	3	2	2	2
Supporting large amounts of data	3	4	3	1	1
Popularity	5	2	4	1	3
Total	**34**	**31**	**24**	**14**	**16**

Decision tree gave the most suitable results in terms of good patterns when small dataset provided, follow by association rules. Clustering also able to identify the fraud outliers and each outlier covers few cases but not better than decision tree and association. Naïve bayes and neural network gave worst results among all. Association took minimum time to produce a finalized model; decision tree took average time while remaining three algorithms took much more time because of data preparation. Decision tree and association was easy to implement while remaining algorithms are much difficult, because they all need normalized data. We first got understanding of normalization, select min-max method, transformed data into normalized view and then implement them. Decision tree is very easy to understand, even by non technical person. Association rules are like if-else conditions, by giving some efforts association can be presented to other people. In clustering, presentation of clusters is not very difficult to understand but in detail level of each cluster, it is a bit complex. Naïve bayes and neural network have complex presentation.

According to our experiments, we observed that association worked better when we increased size of our dataset. Decision tree and clustering gave just slightly below results but in naïve bayes and neural network, we got slightly better results. Decision tree, association and clustering produced better results on having noisy data, while remaining algorithm did not. Decision tree has most simple output presentation follow by association, but remaining three algorithms have not any simplicity as far as output concerns. Association gave better on large size of dataset. Decision tree and clustering gave just slightly below results but in naïve bayes and neural network, we got slightly better results. Figure 10 shows the survey usage of different data mining algorithms. According to the survey, decision tree is the most popular data mining algorithm compare to any other algorithms. If we pick only 5 algorithms those we used, then clustering is second most popular algorithm followed by neural network, association and naïve bayes.

H. Survey Results Analysis

We have conducted a survey according to our work and we got the following results.

Comparison Criteria	**Data Mining Techniques**				
	Decision Tree	Association	Clustering	Naïve Bayes	Neural Network
Training Volume (Low size) against Patterns Level	3	4	2	2	3
Model creation Time	4	4	4	3	2
Ease of implementation	5	3	4	3	3
Ease of Presentation	5	3	3	4	5
Extensibility	5	3	4	3	1
Efficiency	5	5	4	3	3
Simplicity	5	4	3	2	3
Supporting large amounts of data	5	3	2	4	4
Popularity	5	4	3	2	3
Total	**42**	**33**	**29**	**26**	**27**

Raking in each of the question is slightly different compare to our ranking and total ranking points are also not same as we have got from our experiments but if we compare their ranking with our ranking. It gave us the same results and same ranking that we have. Selection of answers on each survey question is based on maximum frequency. For example, If we have 'popularity' as a question and we got 20 votes of '5', 14 votes '4' and 8 votes of '3' in decision tree than we picked '5'. Figure 11 show the bar graph of the user survey results, which is expressing the ranking of the techniques, which we consider in our study.

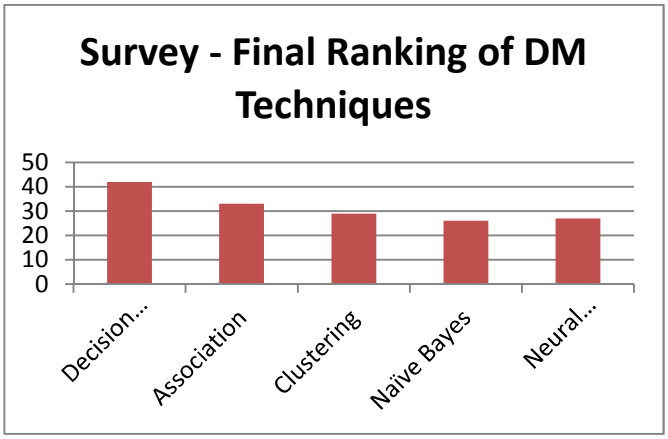


Fig. 11 – Bar graph of survey results for Data mining comparison

This survey gave us motivation and inspiration regarding our work and our direction

I. Final Ranking of Algorithms

On the basis of experiments and scoring of comparative analysis, we concluded the following ranking of different data mining algorithms:

1. Decision Tree
2. Association
3. Clustering
4. Artificial Neural Network
5. Naïve Bayes

VII. FUTURE WORK

In this study, we have tried to compare five different data mining algorithms. By our extensive efforts and research, we are able to find the most suitable algorithm(s) of data mining for fraud detection on the area of branchless banking. As a future work, comparison among these five algorithms can be tried according to different comparison factors excluding our factors. Another future work, further algorithms which we did not include in our study, can be used to compare and find out, whether they are more efficient, reliable and most suitable techniques for fraud detection on the area of branchless banking.

VIII. CONCLUSION

Branchless Banking is growing very rapidly in financial sector. It provides different services in term of transactions. Different transactions can be performed in branchless banking on different channel like Mobile, Agent without the involvement of bank. But on the other hand there are some security laps in branchless banking and there is a need to eliminate or reduce these laps in order to secure customers as well as bank. Primary responsibility of branchless Bank is to secure and satisfies the customer from any kind of frauds and crimes. Crimes include an open range of fraud and illegal activities. It has impact on branchless banks in several areas including financial, operational, and psychological. In this paper, we have learned and analysed different data mining techniques like decision tree, clustering, association rules, naïve bayes, neural network and then we have applied these different techniques on our data of branchless bank. After getting results from all different techniques on certain comparison criteria, we have found out and suggested that "decision tree" is the most suitable technique for fraud detection on branchless banking followed by "association rules" as second best suitable technique. We have also found the comparison factor by which we compared the results and recommend the most suitable technique.

Reference

[1] Adrian Gepp, J. Holton Wilson, Kuldeep Kumar, Sukanto Bhattacharya, A Comparative Analysis of Decision Trees via Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection, Journal of Data Science, 2012

[2] Tom Fawcett, Foster Provost, Adaptive Fraud Detection, Journal of Data Mining and Knowledge Discovery, 1997

[3] Kate Smith,Clifton Phua, Vincent Lee, Ross Gayler, A Comprehensive Survey of Data Mining-based Fraud Detection Research, Journal of CoRR, 2010

[4] Andrew Fast, Lisa Friedland, Marc Maier, Brian Taylor, David Jensen, Henry G. Goldberg, John Komoroske, Relational Data Pre-Processing Techniques for Improved Securities Fraud Detection, Conf. on Knowledge Discovery and Data Mining, ACM, 2007

[5] Aihua Shen, Rencheng Tong, Yaochen Deng, Application of Classification Models on Credit Card Fraud Detection, 2007

[6] Hian Chye Koh, Gerald Tan, Data Mining Applications in Healthcare, Journal of Healthcare Information Management, Volume 19, No.2, 2005

[7] Xiaowei Ying, XintaoWu, Daniel Barbará, Spectrum Based Fraud Detection in Social Networks, Conf. on Computer and Communication Security, ACM, 2010

[8] WEKA (Waikato Environment for Knowledge Analysis), "http://www.cs.waikato.ac.nz/ml/weka/", last accessed: 14 July 2012

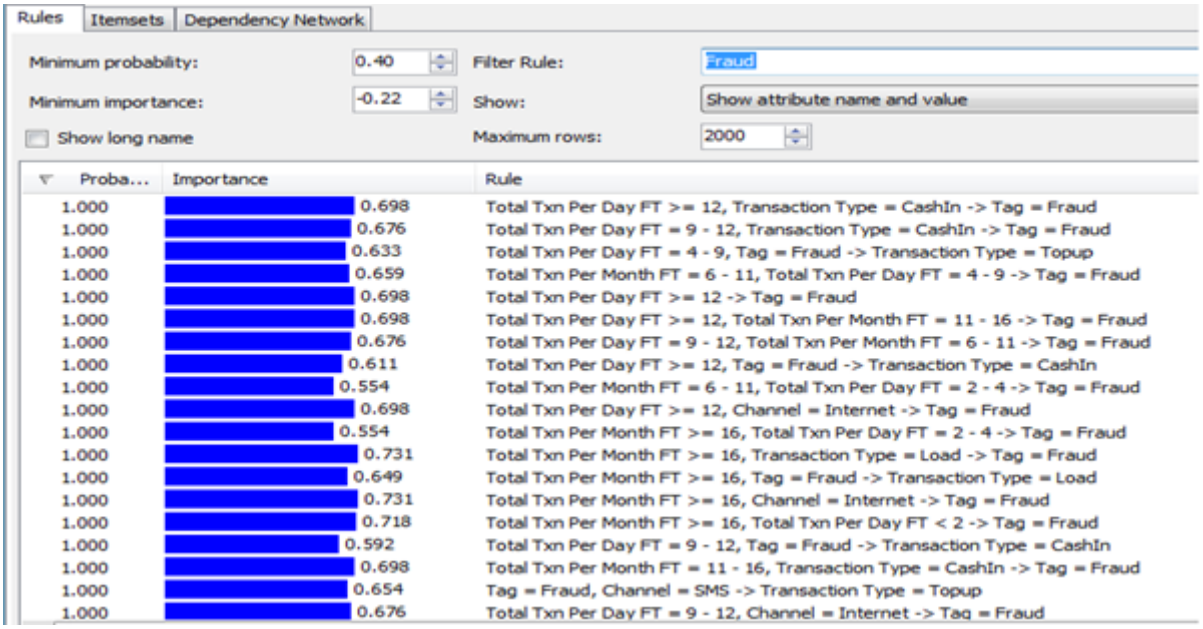
[9] M Afshar Alam, Sapna Jain,Ranjit, Comparison And Evaluation of Scaled Data Mining Algorithms, International Journal of Computer & Organization Trends, Volume 1, No. 3, 2011

[10] Evaluation Criteria for Data Mining Systems, "http://ebookbrowse.com/evaluation-criteria-for-data-mining-systems-pdf-d230680884", Last accessed: 29 October 2012

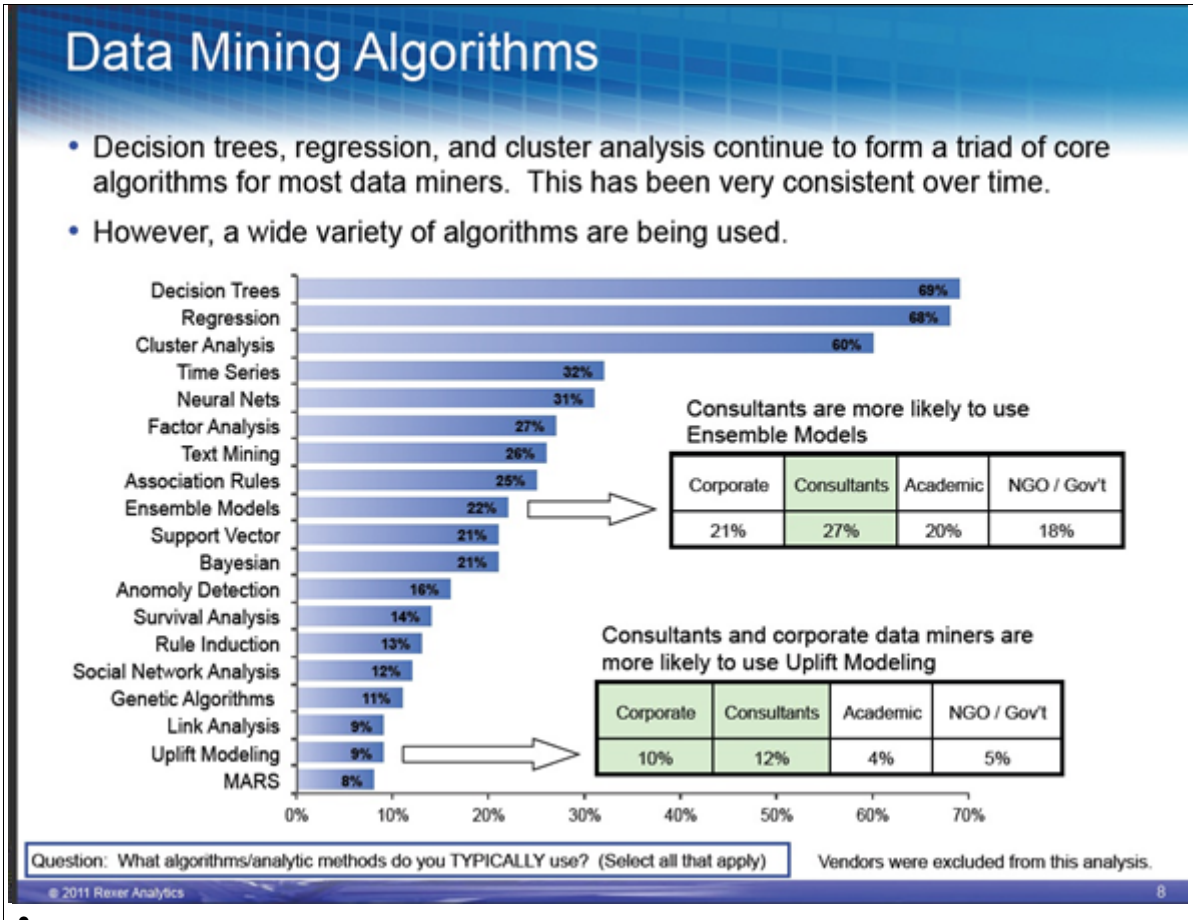
[11] Microsoft SQL Server Analysis Services 2008, http://www.microsoft.com/sqlserver/en/us/solutions-technologies/business-intelligence/analysis.aspx, last accessed: 04 November 2012

[12] M. Talha Umair, Dr. Saif ur Rahman, Fraud Detection Using Data Mining Technique on Branchless Banking, not published, 2012

APPENDIX A



- Fig. 2 - Apply filter (Regular expression) on association rules and get fraud case rules only.



- Fig. 10 - Survey Analysis (Popularity of data mining algorithm according to usage)

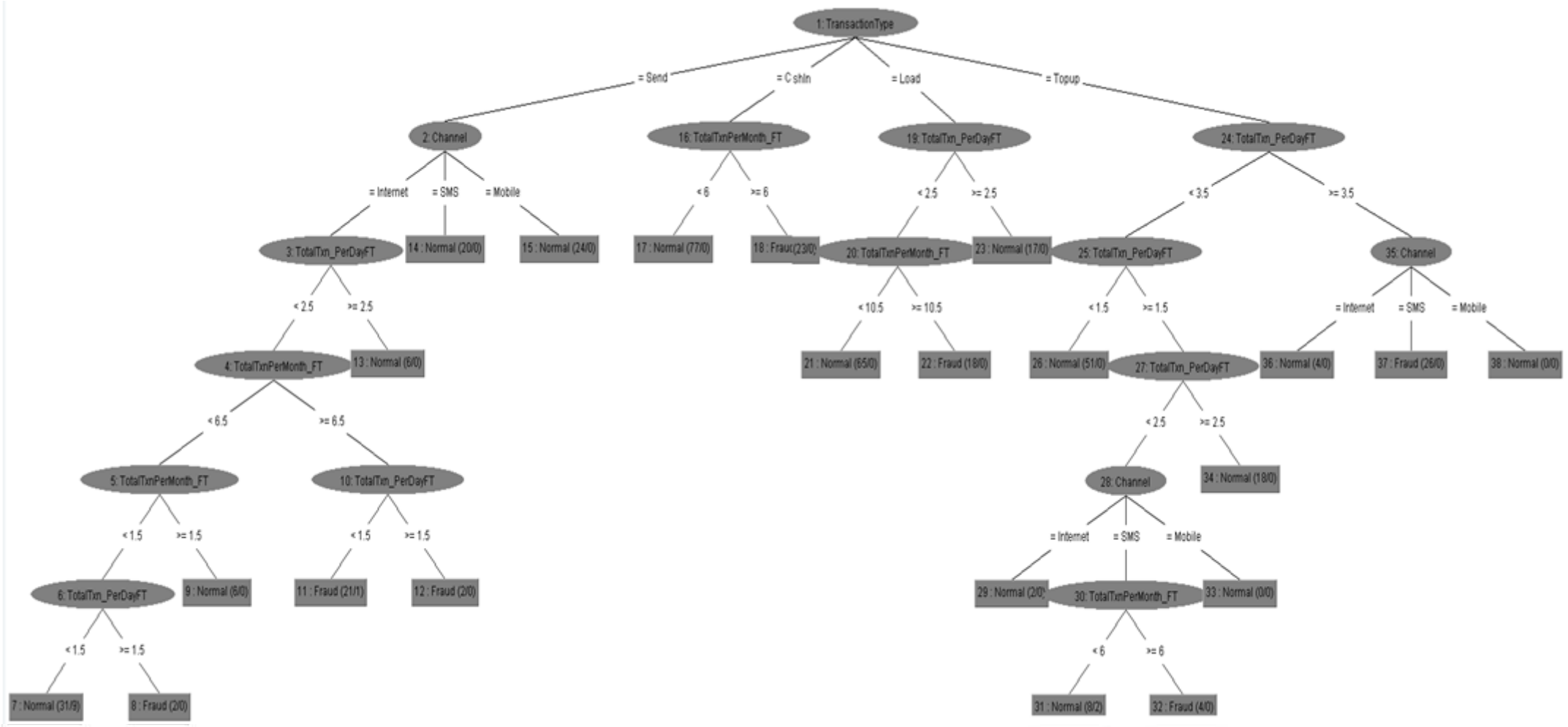


Figure 9 – Decision Tree using Random Tree Algorithm