A Review of Trends and Opportunities for Data Mining Applications in Telecommunications Industry

Parmpreet Kaur[1] and Jyoteesh Malhotra[2]

[1,2]*Department of Computer Sc. & Engineering,*
GNDU Regional Campus, Jalandhar, Punjab, India
E-mail: [1]*parmpreet89@gmail.com,* [2]*jyoteesh@gmail.com*

Abstract—**The development of mobile networks and internet technologies have created a lot of pressure on the telecommunication industry. As a result telecommunication companies are working in highly challenging and competitive environment nowadays. Large amount of data is generated from various systems and this data is used for solving business problems that need urgent problem solving and handling. These data include call detail data, customer data and network data. Data Mining technology and business intelligence (BI) techniques are widely used for handling the business related problems. The key application areas of Data Mining in telecommunication industry are fraud detection, network fault isolation and improving market effectiveness. The aim of this paper is to explore various data mining tools and techniques and to check how they can be used to detect telecommunication fraud, fault and improve market effectiveness.**

Keywords: Data Mining, Telecommunications, Fraud Detection, Network Fault Isolation, Marketing & Churn

I. INTRODUCTION

The evolution of technology has enabled collection and storage of huge amounts of data. The size of databases today can be very large and can range up to terabytes. Practically it is impossible to analyze such large volumes of data using traditional techniques. For this reason data mining has gained a lot of attention. We can find and explore data, generate results, and learn from data. [1]

Data mining is the process of automatically finding and extracting useful information in large amount of data repositories. It is supported by:

1. Data availability.
2. Affordable processing power.
3. Inexpensive data storage.
4. Many commercial data mining tools are available.

Applying data mining to any industry depends on following two factors: the data that are available and the business problems faced by the industry. Data mining can be applied to any business, but in this paper we will describe how it can be used in the telecommunications industry. Telecommunications industry is data-dependent. Telecom companies have the detail records of the customers and the calls made by customers. [2]

The data mining process involves pre-processing of data, analyzing the data, post-processing of the data and the evaluation of results. For each of these steps some method is chosen based on the requirements and at the end of the process useful data is converted into information. This is depicted in the diagram below.
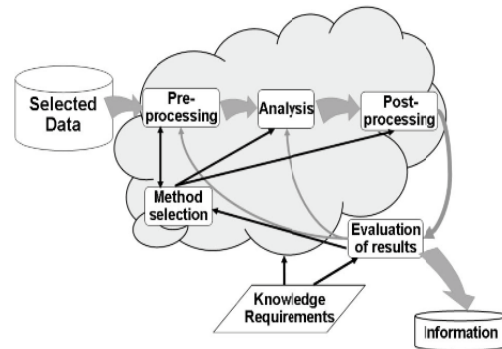

Fig. 1 Structure of Data Mining Process

A telecommunications network provides mobile services for a large geographical area. This area can cover, for example, a whole country. A mobile-phone user can move around the area without losing his connection to the network. This is achieved by placing Base Transceivers Stations (BTS) to give continuous coverage all over the area. A BTS has one or more transmitter-receiver pairs called Transceivers (TRX). Transceivers beam through BTS antennas, which are pointed to cover an area called a cell. When a mobile phone makes a call, it creates a connection to a BTS. From the BTS the call is typically forwarded to a transmission network, which connects all the BTSs to other network elements and to outside networks like Public Switched Telephone Networks (PSTN). When the match occurs in any of the BTS, the call is connected. [4]

This paper is organized as follows. First of all we describe how data mining can be used in telecommunications. Next we describe various types of telecommunication data. Next we explain various applications of data mining in telecommunications i.e. fraud detection, network isolation and marketing. Then we specify the work dome in this field. Next section specifies various issues and challenges and future trends. Final section gives a conclusion of the paper.

Telecommunications network management requires rapid decision-making, which can be supported by data mining methods. The decision-making is based on

information extracted from large amounts of data that are continuously collected from networks. [4]

For a telecommunication organization to achieve dominance in market, following three strategies are open [7]:

1. They must excel at being the best low cost provider of services, making them the market dominator.
2. They be superior at quality of what they are doing, distinguishing them as best provider of the service.
3. They must become best marketers of service by pursuing "customer intimacy" i.e. being a company that responds to customer needs and wants better than anyone else.

II.   II. TYPES OF TELECOMMUNICATION DATA

The first step in the process of data mining is to understand the data. Without understanding, development of useful applications may not be possible. In this section we describe the three main types of telecommunication data. If the raw data is not suitable for data mining, then transformation of data is done to generate data that can be mined. [3]

A.   Call Detail Data

Whenever a call is made on a network, the information about the call is saved as a call record. The call record contains information which describes the important features of every call. Such record usually consists of the originating and terminating phone numbers, the date and time of call and the duration of the call. The call record detail of every customer must be put into a single record that describes the customer's calling behaviour for extracting useful knowledge. This helps in generating customer profiles which can be used for mining data for marketing purposes. [5] Information that can be obtained from the call detail record include

1. Average call duration.
2. Average number of call received per day.
3. Average number of call originated per day.
4. Percentage of no-answer calls.
5. Percentage of weekday calls
   (Monday – Friday).
6. Percentage of day time calls (office hours)[6]

B.   Network Data

A Telecommunication network consists of different types of equipment, which are made of many interconnected components. Each of these components can generate a status and error message that generates huge quantity of network data. The data is normally stored and analyzed so as to support network management functions such as fault isolation and detection. Data mining technology helps to perform these functions by automatically extracting knowledge from the network data. [5]

C.   Customer Data

Telecommunication industries maintain a huge database of information of all their customers. The information consists of names, address, service plan, contract information, credit amount and payment history. These data are often used along with other data for example using customer data along with call detail data to identify phone frauds. Information about the customer can include [6]:

1. Name of the customer.
2. Address of the customer.
3. email id, additional contact nos.
4. Payment history.
5. Service plan etc.

III.   APPLICATIONS OF DATA MINING IN TELECOMMUNICATIONS

The main factors on which Data Mining applications depend include the details of the problem to be solved by the Data Mining and the availability of Data for mining. The main reason for the importance of Data Mining applications in the Telecommunications industry is the availability of large volumes of data. [6] There are many prospects of data mining techniques in telecommunication which may include predicting which customers are likely to default on payment, catching fraudulent activities, identifying of telecommunication patterns, improving resource utilization and service quality and allowing multi-dimensional data analysis to enhance the understanding of customer behaviour. Information obtained from data mining techniques can be used for application like market analysis, fraud detection, science exploration and retention of customer to production control. Data mining helps to reduce company's losses by making good predictions about the possible business outcomes. [5]

There are three major applications of data mining in telecommunication industry. These are:
1. Fraud Detection.
2. Network Fault Isolation & Prediction.
3. Marketing and customer relationship management (CRM).

A.   Fraud Detection

Fraud is very serious issue faced by the telecommunication industry since it leads to the loss of revenue which may range to billions of dollars. As defined by Gosset & Hyland 1999, the telecommunication fraud can be stated as —any activity of using the telecommunication service intention of paying the company providing the service. [6] Fraud detection is important to the telecom industry because companies providing telecommunications services lose a significant amount of revenue due to frauds. [8] In order to identify fraud, data mining application can be used to analyze large amount of cellular call data which

can be used to generate possible patterns. These patterns specify a customer's behaviour with respect to some pattern of fraud. The monitors are then fed into a neural network that determines when there is an evidence of fraud to raise an alert. Data mining also helps in detecting fraud by identifying and storing the phone numbers known to be used fraudulently. [5]

Telecommunication fraud can be classified into two main types Subscription fraud & Superimposition fraud. When a customer opens an account with the intention of never paying it is said to be Subscription fraud and when a unauthorized person gains illegal access to the account of a legitimate customer Superimposition fraud is said to occur. Telecommunication companies consider that superimposition frauds are the most significant problems. Both subscriptions fraud and Superimposition fraud should be detected as soon as possible and customer account should be deactivated immediately. [6]

Customer data can also be used for fraud detection. For example credit information and price plan can be in used in fraud analysis. Another method commonly for fraud detection is to create profile of customer's calling behaviour and compare activity against this behaviour. This behavior can be generated by defining the call detail records for a particular customer. Fraud can be detected immediately after it happens if the call details records for that customer are updated regularly. Fraud detection system work at customer level, not for individual call. Fraud detection involves predicting a rare event where the class distributions involved is complex. [6] Hence, data mining can be used to avoid loss revenues of telecommunication operator due to fraud.

B. Data Mining Techniques for Fraud Detection

The following data mining techniques can be used for fraud detection:

1) Neural Networks

Neural Networks calculate user profiles in an independent manner, thus adapting easily to the behaviour of the various users. [9] In order to differentiate between legitimate user and fraud, feed-forward neural networks (FF-NN) can be used. The problem involves the need to adapt the profiles so that the input-output mapping corresponds to the input-output pairs provided for that profile. The evaluation of performance is done by a Receiver Operating Characteristic (ROC) curve. ROC curve is graphical representation of the trade off between the true positive rates and the false positive rates for every possible cut off point separating overlapping distributions. [8]

2) Decision Trees

Learning algorithms commonly use divide-and-conquer approach. The input space is divided to maximize information gain or specify expression of knowledge change. This approach leads to tree-like data structures. The aim is to have leaves that contain objects of the same class. In particular, fraud cases are characterized by lower deviation values than normal use, which specifies that fraudsters show some kind of "compact" behaviour. They tend to place long calls. [8] For example consider a decision tree for the weekly representation of the users:

IF MeanCalls<0.86 THEN class=1 confidence: 71.98%, coverage: 70.48%)

IF MeanCalls>0.86 AND StdDur<129.5 THEN class=2 (conf.: 97.5%, cov.: 41.5%).

The first rule says that if a user places less than 1 call per day his is a legitimate user with confidence 72%. According to the second rule, if the mean number of calls in a week is more

that 1 (that is at least 7 costly calls in the week) and the standard deviation of their duration is less than 2 minutes, then the user is a fraudster with confidence 97.5%.

3) Agglomerative Clustering

No matter how well a neural network classifier may have performed, still there is no clue about the features that are actually used in order to achieve its performance. In order to further investigate the problem of appropriate user modeling for fraud detection, the hierarchical agglomerative clustering technique can be applied on the data. The aim is to test whether cases from the same class tend to form clusters and if yes, then under which condition. During hierarchical agglomerative clustering the user does not specify the expected number of clusters k. Instead, the algorithm constructs a tree-like hierarchy, a dendrogram, which implicitly contains all values of k. The root of the tree structure defines a cluster that contains all data, while its leafs represent the n clusters, each one containing one of the n objects. The agglomerative clustering algorithm starts with each object representing a cluster, called a singleton, and proceeds by fusing the closest ones until a single cluster is obtained. Therefore, a measure of dissimilarity between two clusters must be defined. Two different distance measures, namely the Euclidean distance and the correlation between objects, are used. Clustering quality can be judged by means of appropriate statistics such as the agglomerative coefficient (AC) and the cophenetic coefficient (CC). [8]

C. Network Fault Isolation & Prediction

Telecommunication networks consist of complex configurations of software and hardware. Since the

industry requires network efficiency and reliability, most of the network elements are designed to generate status and alarm messages in case of any problem. Expert systems were designed to handle such alarms. Network fault isolation in the telecommunication industry is a quiet difficult task because of the following reasons. Huge volumes of data are available and a single fault can generate different alarms which may not be related with each other. Hence alarm correlation is very important in predicting network faults. An active and rapid response is very much essential for maintaining the reliability of the network. Data mining techniques like neural networks, classification and sequence analysis can be used for identifying faults. The telecommunication Alarm Sequence Analysis (TASA) is a Data Mining tool which provides fault identification by identifying recurrent patterns in algorithms. It can be used to generate an alarm correlation system, which can be used to identify faults. Genetic algorithms can also be used to predict telecommunication failures. Time weaver is such a genetic algorithm which has the capability to operate directly on the raw network data. Standard tools of classification can be used to predict future failures but it has drawbacks like, some of the data may be lost in process. [8]

D. Marketing

Telecommunication industries maintain huge amount of information about their customers. Hence, they can use data mining to identify and retain customers and maximize the profit from each customer. [7] Data mining also helps in designing customer's profiles from call record details and then mining these profiles for marketing. The emphasis of marketing in telecom has moved from identifying new customers to measuring customer value and then taking necessary steps to retain profitable customers. [5] Numerous Data Mining techniques can be used to generate customer life time value. To estimate the life time value of a customer we need to estimate how long they will remain with current network. It will help in predicting when a customer is likely to leave and then taking necessary steps to retain the customer. One serious issue that the telecom industries face is customer churn.

Predicting churn, i.e. to check if a customer may leave for a competitor, is an important application of analyzing customer behavior. It is expensive to get new customers then to retain existing ones. Correctly predicting if a customer is about to churn and then convincing him to stay can increase the revenue of a company. [10] Customer churn is used to specify the movement of customer from one provider to another, and 'churn management' describes an operator's process to retain exiting profitable customers. [12]

Telecommunication industry is a large market and is aware of the importance of Customer Management and

Relationship and the impact of Churn. Because of this, developments on following fields are achieved: [11]

Cross Selling and up-selling: maximize profits from existing customers.

Retaining and up-selling: retaining profitable customers or get rid of inappropriate customers in the company profile.

Poaching: to poach (get) new Customers from rival companies.

Obtaining new customers is relatively expensive than retaining existing customers. It is for this reason that telecommunication companies realize that keeping existing customers is more important and churn analysis is an important data mining application areas.

A Person who Churns from one network to another is called a Churner. Two categories of churners are voluntary and involuntary churners: These are explained below [13]:

1. *Involuntary churners* are customers that the company decides to remove from its list of subscribers. This category includes people that are removed for non-payment (customers with credit problem), frauds (customers who cheat), and under-utilization (customers who don't use the phone).

2. *Voluntary churn* occurs when the customer terminates the service. We recognize two types of voluntary churn: incidental and deliberate churn.

Incidental churn occurs, not because something happened in customers lives and not because the customers planned on it. For example: change in financial condition churn, change in location churn, etc.

Deliberate churn occurs due to reasons of technology (customers wanting newer and better technology), economics (price and charges), service quality, social or psychological factors, and convenience reasons. Deliberate churn is the problem that most companies try to solve.

Post paid and Prepaid Churn When considering Post paid churn, the deactivation date, i.e. the date when the customer is disconnected from the network, is equal to the churn date. Because, this is the date when a customer stops using the operator's services. Whereas prepaid segment does not have contract between users and telecom operator, so the definition of Prepaid churn is not simple. However in Prepaid churn, the deactivation date may not match the churn date. [13] Generally it takes a long period of time before a Prepaid customer is disconnected from network. In many cases customers may have churned long before they are disconnected from network. This is the reason why the deactivation date is not suitable indicator for prepaid churn.

Following steps can be followed for the churn prediction [14]:
1. Initially, for each attribute, a threshold value is assigned.

2. The attribute values of the training dataset are compared with the attribute's threshold to declare that a customer will churn or not. Simple if…then …else rules are applied in this process.

3. A model is then constructed for the training dataset.

4. The model is then applied on the test dataset and the results are listed.

5. The above steps can be repeated by varying the threshold values of the attributes selected.

Churn prediction data mining assessment methodology
The purpose of this research is to assess the performance of various data mining techniques when used in context of churn prediction. The methodology consists of three parts [12]:

1. An IT infrastructure, which includes a common customer base, attributes and transactions, modeling parameters, model results, etc.

2. A model-independent knowledge discovery procedure to discover customer behaviour prior to churn, by using data mining techniques.

3. A set of measurements to measure the performance of models developed by different modeling tools, such as decision tree and neural network.

IV. RELATED WORK

In 2006, Shin-Yuan Hung [12] studied how data mining can be applied to telecommunication for churner predication and what measures need to be undertaken to ensure that churn does not occurs.

Sen Wu, Naidong Kang [2] in 2007 stated that customer data available should be used by the operator to analyze the common characteristics of fraudulent behaviour of customers in telecom industry systematically for detection frauds.

D. Camilovic [1] in February 2008 emphasised that CRM and churn cannot be practiced in business without tracking patterns within customer data. It highlights the importance of using the data available to predict churn.

Gary M. Weiss [3] in 2009 explained various types of telecom data and how data mining can be used in applications like marketing (for CRM and Churn), fraud detection and network isolation and for future predictions.

Rob Mattison [7] in his book explains what data mining is and how it can be used for marketing in telecommunications for Churn management.

Kimmo H¨at¨onen [4] in 2009 in his thesis explained how decision making can be implemented on various levels for making decisions and identifies the data mining tools that can be used for this purpose.

Umman Tugba Simsek Gürsoy [11] in 2010 provides an analysis of churn and its impact on the operator and explains what techniques should be used to predict and prevent it.

Frank Eichinger [10] explains what churn is and how we can predict churn in a network and what techniques to use to avoid churn.

Goran Kraljevi´c [13] in 2010 explained how prepaid churn analysis is difficult as compared to postpaid churn and how to deal with prepaid churn using data mining tools and techniques.

Isinkaye O. Folasade [5] in 2011 explores different data mining tools and applications and how they can be used to detect telecommunication fraud, fault and improve market effectiveness.

Constantinos S. Hilas [8] in March 2012 explains how data mining can be applied over data available to an operator for checking abnormal data patterns which may be helpful in detecting fraudulent customers and data.

V. Umayaparvathi [14] in March 2012 explained in her journal what techniques to use for predicting churn in telecommunication environment and what techniques to use to ensure that a customer does not leaves the operator.

Madhuri V. Joseph [6] in February 2013 explained that telecommunication industries have large amount of data(call detail data, customer data and network data) which can be used for solving many business problems that require urgent handling.

Anita B. Desai [9] in 2013 in her journal explains how data mining techniques like neural networks, decision trees and agglomerative clustering can be used for detecting frauds.

V. FUTURE TRENDS AND CHALLENGES

Data Mining has an important role because the data is the primary concern in telecommunication industry. The telecommunication data is mainly in the form of transactions or events and is not at a level for mining so complex pre-processing of data is required. Scalability is an important issue because the industry is handles very large databases. Fraud detection and network fault isolation in the telecom increase the importance of real-time operation. Data Mining should consider the privacy issues. This is so because telecommunication companies maintain private information like whom each customer calls. Another issue with telecom data and its applications involves rarity. Predicting and identifying rare events like fraud and equipment failure is quite difficult for many data mining algorithms and therefore this issue must be handled carefully in order to ensure good results. It is certain that the new Data Mining applications will be developed and deployed which will help to eliminate some of the problems faced by current applications.

VI. CONCLUSION

Data mining applications play an important role in the telecom industry due to the presence of large amount of data and the rigorous competition faced in the sector. The primary application areas are marketing

and Customer Relationship Management, Fraud detection, Network Management. The latest developments in the Data Mining involve enhancement and implementation of existing techniques and methods to ensure continuous growth of telecommunication companies that using them. This paper explains how data mining tools and techniques can be used by telecommunication companies to discover and extract useful patterns from large volumes of data so as to find observable patterns, which can help in identifying and catching fraudulent activities, improving resource utilization and service quality, facilitating multi-dimensional data analysis to improve the understanding of customer behaviour.

ACKNOWLEDGMENT

REFERENCES

[1] D. Camilovic, "Data Mining And CRM In Telecommunications" Serbian Journal of Management 3 (1), February 2008.
[2] Sen Wu, Naidong Kang, Liu "Yang Fraudulent Behavior Forecast In Telecom Industry Based On Data Mining Technology" Communications of the IIMA 1, Volume 7 Issue 4 (3), 2007.
[3] Gary M. Weiss, "Data Mining In The Telecommunications Industry", IGI Global, 2009.
[4] Kimmo H¨at¨onen, "Data Mining For Telecommunications Network Log Analysis", Helsinki University Printing House, 2009.
[5] Isinkaye O. Folasade, "Computational Intelligence In Data Mining And Prospects In Telecommunication Industry", Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS) 2 (4) (ISSN: 2141-7016), 2011.
[6] Madhuri V. Joseph, "Data Mining and Business Intelligence Applications In Telecommunication Industry", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013.
[7] Rob Mattison, "Data Warehousing And Data Mining For Telecommunications", ISBN 0-89006-952-2, Artech House Boston London.
[8] Constantinos S. Hilas, "Data Mining Approaches To Fraud Detection In Telecommunications", 2nd Pan-Hellenic Conference on Electronics and Telecommunications PACET'12, March 2012.
[9] Anita B. Desai, Dr. Ravindra Deshmukh "Data Mining Techniques For Fraud Detection", International Journal of Computer Science and Information Technology (IJCSIT) ISSN: 0975-9646, Volume 4(1), 2013.
[10] Frank Eichinger, Detlef D. Nauck, and Frank Klawonn, "Sequence Mining For Customer Behaviour Predictions In Telecommunications".
[11] Umman Tugba Simsek Gürsoy, (2010) "Customer Churn Analysis In Telecommunication Sector", Istanbul University Journal of the School of Business Administration, ISSN: 1303-1732, Vol:39, No:1[12] Shin-Yuan Hung, David C. Yen, Hsiu-Yu Wang, "Applying Data Mining To Telecom Churn Management", Expert Systems with Applications, 2006.
[12] Goran Kraljevi´c, Sven Gotovac, "Modeling Data Mining Applications For Prediction Of Prepaid Churn In Telecommunication Services", AUTOMATIKA 51, ISSN 0005-1144, 2010.
[13] V. Umayaparvathi, K. Iyakutti, "Applications Of Data Mining Techniques In Telecom Churn Prediction", International Journal of Computer Applications (0975 – 8887) Volume 42– No.20, March 2012.