Data Mining in Telecommunication:A Review

Suman H. Pal, Jignasa N. Patel
Department of Information Technology,
Shri S'ad Vidya Mandal Institute of Technology

Abstract **—Telecommunication is one of the first industries to affect data mining technology. These companies have different types of customer are there. When companies save this data and companies have huge amount of data .In which different types of data are included call-detail data, customer data, and Network data. Companies handle this types of large amount of data .Telecommunication companies are used in data mining technology. This paper describes data mining in telecommunication and its application help Telecommunication Industry. Using data mining technology for churn prediction in Telecommunication industry. Telecommunication companies face impact of companies because some of the customer who is at risks of leaving a company the proposed model is possible to predicting customer churn behavior in advance. This paper describes how data mining and its application helps telecommunication industry in various aspects. . In this paper, an unbalanced data set is characterized by an uneven class distribution where the amount of fraudulent instances is substantially smaller than the amount normal instances. This will result in a classifier which is most likely to classify data has belonging to the normal class then to the fraud class.**

Index Terms— **Data Mining; Churn Prediction; Application; Telecommunication; Fraud detection**

I. Introduction

The Telecommunication industry generates tremendous amount of data .In which different types of data included call detail data that is describe calling behavior of that customer that is connected from Network that is nothing but network data .human based expert system was handle the these large amount of data [1] .information from this data automatic or semi automatic method should be used .this system helped to detect the frauds and network issues. but this system is time consuming system as it involves much of knowledge from human expert That's way this method are not used .At that time the data mining was new. so Telecommunication companies started with this technology .data mining are also adopt the large amount of data .they first to adopt data mining technology in telecommunication. But telecommunication pose different issues for data mining .1.The first issues is Telecommunication data may contain tremendous amount of data (record) 2.second is raw data is not always suitable for data mining 3.last issue is concerned with real time performance of data mining application that is fraud detection .when we deal with these issues then efficiency of data mining increase 100%.Telecommunication industry are also increasing customer churn behavior . When the customer creates an unnecessary and undesired financial problem on the company and ultimate may lead to sickness of the company, detecting the customer.

II. Data Mining

The Meaning of data mining is extracting knowledge from large amount of database and data .it is the one part of (KDD process) Knowledge discovery in database .the actual data mining task is the automatics and semi automatics analysis of large amount of data to extract previously unknown interesting pattern generated such as group of data record ,useless record and association rule mining. This is involve using data base technique [2]. Data mining technique are applied in telecom data base for various purpose .each user different type of telecom data depending on the purpose. the data generated by telecom industries are broadly grouped into 3 types 1.customer data2.network data 3.bill data.

III. Data Mining in Telecommunication

Data Mining is applied for Telecommunication in various factors [2].

A. *Churn prediction:*

Statement of customer who are at risk of leaving a company is known as churn prediction in telecommunication .the company should focus on customer and make different effort retain them. this application is very important because it is less expensive to retain a customer than acquire a new.

B. *Insolvency prediction :*

Increasing due to bills becoming crisis problem for any telecommunication company because of the high competition in telecommunication market, companies cannot afford the cost of insolvency. to detect this problem to used data mining technique .

C. *Fraud detection:*

Fraud is very expensive activity for the telecommunication industry, therefore companies should be try to identify dishonest user and their usage technique.

IV. Churn Prediction in Telecommunication

In CRM(customer relationship management in telecommunication companies is the ease with which customer can move to a competitor ,a process is known as “churning” Churning is an expensive process for the company ,as it is much cheaper to retain a customer than other[4]. In many place statics method have been applied for churn statement. the main objectives of the application to be presented here to find out which type of customer of a telecommunications company is likely to churn example is BSNL[3] . BSNL is the telecom companies suffer from churning customer. it provides many services like the internet ,fax, post and prepaid mobile phones .The churn prediction problem represented three phases 1. Training phase 2. test phase 3.prediction phase. the input for this problem included the data on past calls for each mobile .for the training phase, labels are provided in the form of a list of churners. the model is trained with highest accuracy.

V. Need of Data Mining in Telecommunication

Data is the base of telecommunication so data mining is used to performed operation on data and get the proper result some of the reasons to used data mining are as follows[2].

A. To Detect Frauds

. Fraud is very dangerous problem for Telecommunication companies, resulting in billions of dollars of lost.

B. To Retain Customers

By providing the data mining tool we can learn and research on the customers data base .which will useful to know how to deal with customer and how to agree them.

C. To know the customer

By learning the behavior of the customer we can understand them.

D. Product and services which yeild highest amount of profit

As it is very highly database ,so it contains transactions made by customer .That Transactions may be purchased product, and services.

E. Factor that influence customer to call more at certain time

By studying behavior of customer, we can come to know what are the factor that influence customer to call more at certain times.

VI. Application of Data Mining in Telecommunication

Telecommunication companies maintain a highly amount of information about this extremely competitor environment have great motivation for use this information for these reasons the telecommunication Numerous data mining application have been deployed in the application industry .the most application fall into the following three categories: marketing and customer profiling , fraud detection, and network fault isolation and prediction[2].

A. Marketing And Customer profiling

Marketing is an important thing for the telecommunication industry. Telecommunication companies maintain an enormous amount of information about their customer and due to an extremely competitive environment. this information includes his personal information like age, gender, lifestyle, knowledge etc. then the Data mining tools are applied and segmentation is done on this huge amount of data .that is used to built customer profile and then it is helps to build marketing strategies, planning for the future decisions, performance measurement ,result tracking.

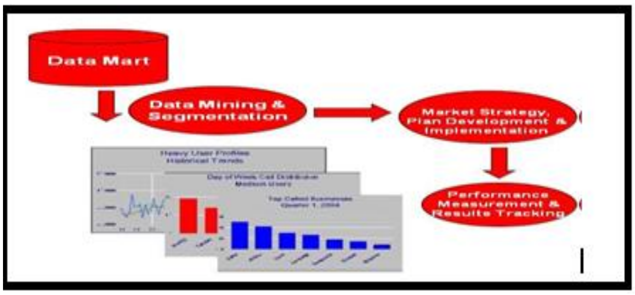
Data Mart
Data Mining & Segmentation
Market Strategy, Plan Development & Implementation
Performance Measurement & Results Tracking
Heavy User Profiles Historical Trends

Fig.1 Marketing/Customer Profiling

B. Fraud Detection

Fraud detection is very biggest problem for telecommunication companies, resulting in billions of lost tax each year. Fraud can be divided into two categories 1.fund fraud 2.superimpostion fraud in fund fraud occurs when a customer open and account with the decide of no paying the account .in superimposition fraud occur when a profit illicit access to the account of a legitimate customer. There are some method for identify fraud that do not involve comparing new behavior against a profit of old behavior .the call details record to summarized to obtain calling behavior if the call detail summarized are modified in real time than the fraud can be detected soon after it occur

.

Fig.2 Fraud Detection

C. Network Fault Isolation and Prediction

Monitoring and maintaining telecommunication in network is an important task ,as these network become increase complex expert system .where developed to handle the alarms generated by the network element, data mining application have been developed to identify and predict network faults.

Fault identification can be quite difficult because a single fault many result in a cascade of alarms many of which are not associated with the root cause of the problems. so in order to identify the network fault the alarms should be analysis automatically so, the data mining helps to automatically analysis and detect the faults so they can resolve.

VII. Types of Telecommunication Data

The first step in the data mining process is to understand the data. Without such an understanding, useful applications cannot be developed. In this section we describe the three main types of telecommunication data. If the raw data is not suitable for data mining, then the transformation steps necessary to generate data that can be mined are also described [2].

A. Call Detail Data

Every time a call is placed on a telecommunications network, descriptive information about the call is saved as a call detail record. The number of call detail records that are generated and stored is huge. Call detail records include sufficient information to describe the important characteristics of each call. Call detail records are generated in real time and therefore will be available almost immediately for data mining.. Call detail records are not used directly for data mining, since the goal of data mining applications is to extract knowledge at the customer level, not at the level of individual phone calls. Below is a list of features that one might use when generating a summary description of a customer based on the calls they originate and receive over some time period P:

- average call duration
- % no-answer calls
- % calls to/from a different area code
- % of weekday calls (Monday – Friday)
- % of daytime calls (9am – 5pm)
- average # calls received per day
- average # calls originated per day
- # unique area codes called during P

B. Network Data

Telecommunication networks are extremely complex configurations of equipment, comprised of thousands of interconnected components. Each network is capable of generating error and status messages, which leads to a tremendous amount of network data. This data must be stored and analyzed in order to support network management functions, such as fault isolation. This data will minimally include a timestamp, a string that uniquely identifies the hardware or software component generating the message and a code that explains why the message is being generated. The call detail data, network data is also generated in real-time as a data stream and must often be summarized in order to be useful for data mining.

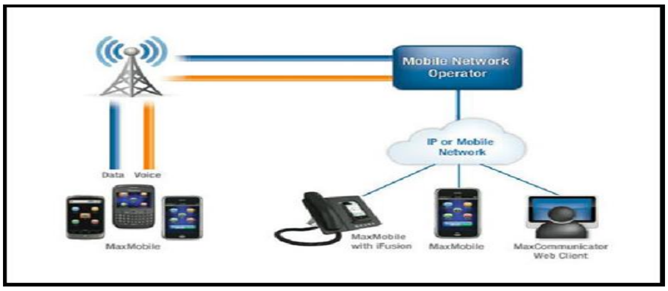

Fig.3 Network Architecture

C. Customer Data

Telecommunication companies, like other large businesses, may have millions of customers. By necessity this means maintaining a database of information on these customers. This information will include name and address information and may include other information such as service plan and contract information, credit score, family income and payment history. This information may be supplemented with data from external sources, such as from credit reporting agencies. Because the customer data maintained by telecommunication companies does not substantially differ from that maintained in most other industries, the applications described do not focus on this source of data. However, customer data is often used in conjunction with other data in order to improve results. For example, customer data is typically used to supplement call detail data when trying to identify phone fraud.

VIII. Data Mining Approaches to Fraud Detection in Telecommunication

Fraud detection is important in telecommunications industry because these companies and suppliers of telecommunications services lose a significant proportion of their revenue as a result. Moreover, the modeling and characterization of users' behavior in telecommunications can be used to improve network security, improve services, provide personalized applications, and optimize the operation of electronic equipment and/or communication protocols[5]. This is to appropriately model user behavior and then apply automated intelligent techniques in order to distinguish normal from fraudulent use. The main are the technical fraud, the contractual fraud, the procedural fraud, and the hacking fraud. The first three usually burden the economics of the service provider, while hacking fraud also harms the subscriber. Hacking fraud is usually met in the form of the superimposed fraud where the fraudster (hacker) uses a service concurrently with the subscriber and burdens his account.

Research in telecommunications fraud detection is mainly motivated by fraudulent activities in mobile technologies. Fraud detection methods can be supervised or unsupervised. This is observed by the following algorithm.

A. Neural Network

In order to test the ability of each profile to discriminate between legitimate employment and fraud, provender -forward neural networks (FF-NN) were used as classifiers [5,6]. The problem is a supervised eruditeness one with the undertaking to adapt the weights so that the stimulant -output mapping corresponds to the stimulus -output twosome the teacher has provided. A detailed presentation of our approach path can be found in Figure 4. The evaluation of each classifier's performance was made by means of the corresponding Receiver Operating Characteristic (ROC) curve. A ROC curve is actually a graphical representation of the trade-off between the true positive and the false positive pace for every possible cut off point that separates two overlap distributions.

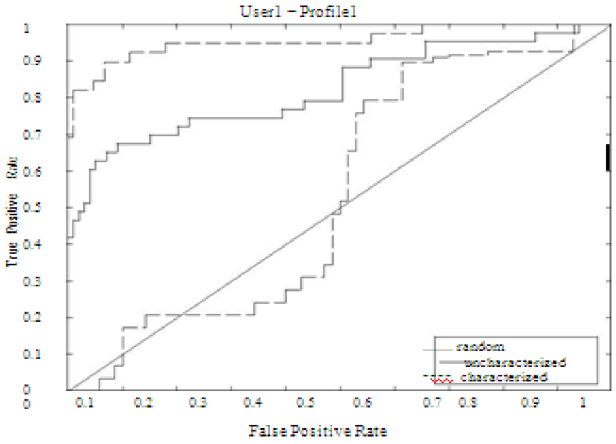

Fig 4. Example of ROC curves

B. Decision tress

Learning algorithms usually use a divide-and-conquer approach. The input space is incrementally divided using splits that maximize information gain or some other expression of the change in knowledge. This approach leads to tree-like data structures [5]. The aim is to have leaves that are as pure as possible, i.e. contain objects of the same class. Appropriate purity measures should be used and the procedure will ideally lead to pure leafs. As the most common measure of this purity one uses the Kullback - Leibler distance, or relative entropy. This is used to express the information of the parent node minus the information of any possible division.

In Figure 5 an example of the construction of rules for the distinction between normal and fraudulent use. Its right branch (Mean Calls>0.857) covers 77% of the fraudulent cases. An interesting result is that the standard deviation of calls' duration is important in fraud identification. In particular, fraud cases are characterized by lower standard deviation values than normal use, which implies that fraudsters show some kind of "compact" behavior. An expected aspect of fraudsters' behavior is apparent in the lowest leaves of Fig 5 . They tend to place long calls. The most powerful rules that result are:

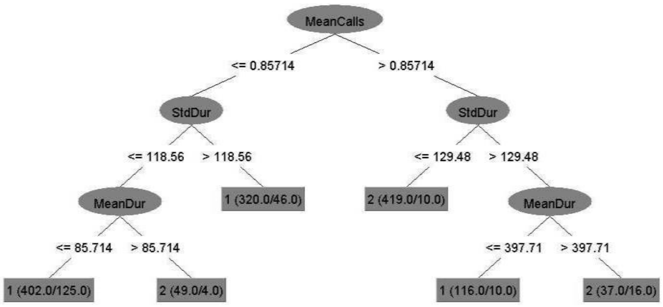

Fig 5 . An Example of Decision Tree for the weekly Representation of the users

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%).

C. Agglometative Clustering

A neural network classifier may have performed, there is no clue about the features it actually used in order to achieve its performance. So, it is difficult to identify important characteristics that led to a successful classification. In order to further investigate the problem of appropriate user modeling towards fraud detection, the hierarchical agglomerative clustering technique shall be applied on the data [5]. The aim is to test whether cases from the same class tend to form clusters and under which condition. During hierarchical agglomerative clustering the user does not specify the expected number of clusters k. 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.

IX. CONCLUSION & FUTURE SCOPE

This paper described how data mining is used in the telecommunications industry. Three main sources of telecommunication data (call detail, network and customer data) were described, as were common data mining applications (fraud, marketing and network fault isolation).

This paper also highlighted several key issues that affect the ability to mine data, and commented on how they impact the data mining process. One central issue is that telecommunication data is often not in a form or at a level suitable for data mining. Other data mining issues that were discussed include the large scale of telecommunication data sets, the need to identify very rare events (e.g., fraud and equipment failures) and the need to operate in real-time (e.g., fraud detection). A more significant issue in the telecommunications industry relates to specific legal restrictions on how data may be used. In the United States, the information that a telecommunications company acquires about their subscribers is referred to as Customer Proprietary Network Information (CPNI) and there are specific restrictions on how this data may be used.

In the case of customers who switch to other service providers, the original service provider is prohibited from using the information to try to get the customer back (e.g., by only targeting profitable customers). Furthermore, companies are prohibited from using data from one type of service (e.g., wireless) in order to sell another service (e.g. landline services). Thus, the use of data mining is restricted in that there are many instances in which useful knowledge extracted by the data mining process cannot be legally exploited. Much of the rationale for these prohibitions relates to competition.

Future work will be in the form of credit application fraud detection. Social network analysis is also of great interest and seems like an appealing alternative to statistical approaches or computational intelligence ones.

REFRENCES

[1] Gray M. Weiss, “Data mining in Telecommunication”, Fordham University ,USA,2009.

[2] Mohsin Nadaf & Vidya Kadam, “Data Mining in Telecommunication”,ISSN:2319-2526,vol-2,2013.

[3] Rahul J.Jadhav,Usharani T.Pawar, “Churn prediction in Telecommunication using Data mining Technology”,IJACSA,vol 2,No.2,Febuary 2011.

[4] V.Umayaparvathi,K.lyakutti, “Application of Data mining Technique in telecom churn prediction”,IJCA(0975-8887)vol 42,No.20,march 2012.

[5] Constantinos S. Hilas, “Data mining approaches to fraud detection in telecommunication”,PACET 12,march 16-18,2012.

[6] Yufeng Kou ,Chang-Tien Lu,Yo-ping Huang, “Survey of Fraud Detection Technique”,IEEE Trans. On Networking Sensing &Control,pp.749-754,2004.