ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Customer Profiling and Anomaly Detection in Telecommunication Industry

Amita Jajoo, Vishnupriya G S, B. Niket

Assistant Professor, Dept. of I.T., D Y Patil College of Engineering, Akurdi, Pune, Maharashtra, India

ABSTRACT: Data telecommunication services are characterized by large number of business systems, numerous data and the use of outmoded means of information acquisition, information to provide in terms of speed, quality, and scope and will greatly lags behind the need of information. Marketing success depends on the study of the preferences of different customer groups, their shopping outlooks and the concept of expense. Based on this condition, customer segmentation can help enterprises to find the proper market positioning for their services. Partitioning can be useful in selecting appropriate customer groups in target market to find the marketing opportunities in current customer groups and to gain competitive advantages through product diversification which is the first stage of proposed system. The next phase focuses on customer churn which aims to identify subscribers who are about to transfer their business to an adversary as price of maintaining a present user (subscriber) is much less than that of registering a new subscriber. The final stage deals with fraud detection whose objective is to target subscription and superimposition deceits.The system is going to analyse the results of all these three phases through WEKA (Waikato Environment for Knowledge Analysis). As compared to the existing system, our proposed system gives superior outcome which includes the analysis of segmentation, churn and uncovers various frauds in one package.

KEYWORDS: Segmentation, churn, fraud detection, data mining.

I. INTRODUCTION

The telecommunication industry plays vital role in message sharing and information passing. In the last few years, a drastic change in the structure of telecommunications companies has been taken place, from public monopolies to private aggregations. The telecommunications industry generates and stores a large amount of data. These data include call detail records, which describes the calls that traverse the telecommunication interconnections, network data, which describes the state of the hardware and software components in the network, and customer statistics, which describes the telecommunication customers. The volume of data is so great that manual analysis of the data is difficult, though not impossible. The need to handle such large amount of data led to the advancement of knowledge-driven expert systems. These automated systems performed significant functions such as identifying fraudulent phone calls and identifying flaws in the network. The problem of this approach is that it is time consuming to obtain the knowledge from human experts (the knowledge acquisition barrier) and, in many cases; the experts do not have the required knowledge. The invention of data mining technology promised solutions to these problems and for this reason the telecommunications industry was an early adopter of data mining technology. Data mining deals with the extraction of previously unknown and potentially useful data which can prove helpful in generating reports and giving precise results.

Telecommunications is one of the most data-comprehensive industries in the world, and a great scope exists for telecom managers to analyse the large volumes of data in order to improve the short-term and long-term operations of their framework. One highly operative tool to aid in this data analysis is the proven process of data mining. Data mining enables the user to view data from many different aspects, categorize the data in new ways and summarize the resulting relationships between seemingly incongruous pieces of data that the software has identified. Rigorous analysis of these relationships can provide managers with the skill to optimize internal network operations and better manage external customer- facing activities such as churn and marketing.

Telecom data about customers such as call detail and customer information can be effectively data mined. Mining these data types can help to identify customer behaviour and determine opportunities to support the goals of expansion of customer base and reduction of customer churn. In face of increasingly competitive market economy, customers

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

have become the most important resources associated with the enterprises development. This evolves the need of segmenting customers into different groups so that the results can be viewed as per the requirement of Telecom Company. This intern helps them to utilize their resources more effectively. Customer churn is a field where mining can be beneficial, as it is becoming ever more important to retain customers and improve wallet share. In the paper, classical data mining technique- decision tree is employed. The term churn is generally used in the telephone industry to refer to all types of customer abrasion whether voluntary or involuntary. Fraud is a consequential issue for telecommunication companies, leading to billions of dollars in loss of revenues each year. Fraud can be divided into two classes: subscription fraud and superimposition fraud. Subscription fraud arises when a customer creates an account with the intention of never paying for the account charges. Superimposition fraud involves an authorized account with some authorized activity, but also includes some superimposed illegitimate activity by a person other than the account holder.

II. RELATED WORK

Data mining is a technique to extract important data from the existing database so that knowledgeable actions can be taken. Customer segmentation based on customer value Customer value has been studied under the name of LTV (Life Time Value), CLV (Customer Lifetime Value), and CE (Customer Equity). The previous researches were done on LTV (Life Time Value) as the sum of the revenues gained from company's customers over the lifetime of transactions after the deduction of the total cost of attracting and servicing customers, taking into consideration the time value of money. Results indicate that boosting of simple classifiers achieves best results and that open-source tools can achieve performance very close to the best solutions. Applicability of various machine learning algorithms that are available within the data mining tool "weka" were studied, to the problem of predicting the churn of telecommunication users.

"Weka" is open-source software that has become standard for the evaluation and application of methods of data mining. "Weka" helps its users to perform the necessary experiments efficiently, as well as enabling the subsequent realization of practical systems for prediction and classification. "Weka" contains an advanced set of tools for data mining and data analysis. It contains implementations of a large number of algorithms. Misuse customer and malpractice customer in the telecommunication area using sequential pattern clustering and genetic algorithm were identified.

III. PROPOSED ALGORITHM

The system is designed for employing the use of company's resources in the most appropriate manner. The aim of the paper is the implementation of three different methods namely, customer profiling or segmentation, predicting customer churn rate which is analysed through decision tree implemented in "weka". It uses a tree-like graph or model of decisions and their possible effects, including chance event outcomes, resource costs, and utility, and finally, detection of frauds or anomalies through comparison with the previously proved deceits.

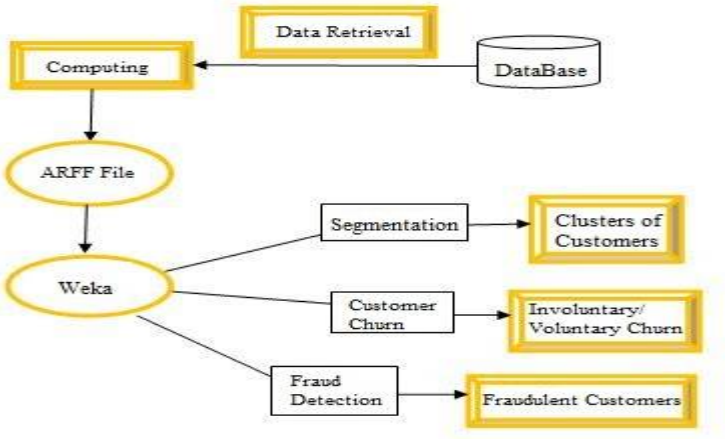

Fig.1.Working process

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

A. SEGMENTATION

Segmentation aims at partitioning customers into clusters that are based on certain criteria so as to analyze the results as per requirement. Segmentation or customer profiling is implemented through k-means algorithm which is a method of vector quantization. Segmentation in the proposed system is based on bill i.e., whether the customer has subscribed to prepaid or postpaid service, zones to which the customer belongs and plan.

Algorithm 1

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The initial assignment of points to clusters can be done at random. In the series of the iterations, the algorithm tries to minimize the sum, over all groups, of the squared within group errors, which are the distances of the points to the respective group means. Convergence is reached when the objective function (i.e., the residual sum-of-squares) cannot be lowered any more. The groups obtained are such that they are geometrically as compact as possible around their respective means. This algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

Formula 1

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the *n* data points from their respective cluster centers.

B. EVALUATING CUSTOMER CHURN RATE

Data sets used for testing purpose in churn prediction contain 500 instances containing both, numerical as well as categorical variables. These instances are reduced by k-means clustering. The input database is compared to historical data with the help of which the churn rate is predicted. The basic algorithm for decision tree is the greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. We are employing greedy strategies because they are efficient and easy to implement.

Algorithm 2

1. Initially the set of selected instances is empty i.e., solution set.
2. At each step
    - Instances will be added in a solution set by using selection function.
    - IF the set would no longer be beneficial
        - Reject items under consideration (and is never consider again).
    - ELSE IF set is still beneficial THEN

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

- Add the current item.

C. DETECTING ANOMALIES

Our schema will identify anomalies through regression algorithm and association rule mapping. Deceits will be found out by comparing with previously proved frauds, called pattern matching. An association rule is an implication expression of the form X Y, where X and Y are disjoint instance sets. Association rule can be measured in terms of Support and Confidence. Support identifies how often a rule is applicable to a given data set, while confidence identifies how frequently items in Y appear in the transactions that contain X. The formal definitions of these matrices are-

Formula 2

$$\text{Support, } s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

IV. RESULTS

This schema generates reports based on customer segmentation, customer churn and fraud detection. For customer profiling it generates a report partitioning customers into different groups based on various different parameters such billing used by customers (Prepaid/Postpaid), zonal classification and plans used. For customer churn, it generates reports predicting voluntary and involuntary churn of telecommunication customers. For anomaly detection it generates reports detecting fraudulent customers using association rule mapping. The schema also shows comparison between existing and proposed system by providing graphs.

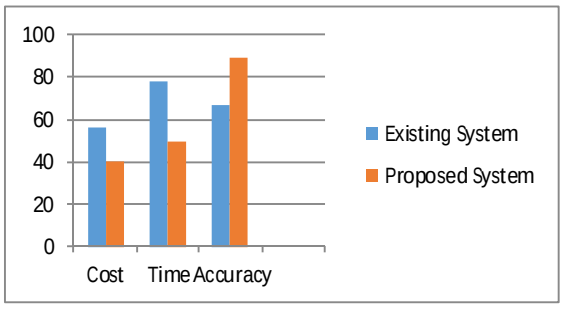

Fig. 2.Comparison of existing and proposed system

It was analyzed that the proposed system showed better statistics in terms of cost and time as compared to the prevailing schema. Our system also proved more accurate and precise while displaying the outcomes of profiling and detecting frauds.

V. CONCLUSION AND FUTURE WORK

In this paper, we optimize the use of resources of Telecommunication Company by grouping customers in different clusters and predicting churn rate using k-means and decision tree algorithm respectively. This is vital for a telecommunication company as it will help them to retain its clients. We also provide an efficient solution to identify misuse and customer using liner regression algorithm and associate rule mapping. This schema serves as a cost effective and time saving solution for a telecommunication company. In future we are going to develop a model for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

business trends and marketing strategy. In this paper we provide effective solutions for customer profiling, predicting customer churn rate and identifying different types of fraud. In future we will develop a complete model for telecommunication industry that will detect all categories of fraud.

REFERENCES

[1]Jinghua Zhao, Wenbo Zhang, Yanwei Liu "Improved k-means cluster algorithm in telecommunications enterprise customer segmentation", Jilin Normal UniversitySiping, China, 2010
[2] Srdjan M. Sladojevic, Dubravko R. Culibrk, Vladimir S. Crnojevic, "Predicting the churn of telecommunication service users using open source data mining tools.", Serbia, Nis, October 5 - 8,2011.
[3] RahmanMansouri, "Application of Data mining in predicting cell phones Subscribers Behavior Employing the Contact pattern", Faculty of Postgraduate Studies Department of Computer University of Najaf Abad Islamic Azad Najaf Abad , Iran, 2010.
[4] Zhangxiao-bin, Gaofeng, Huang hui, " Customer-churn Research Based on Customer Segmentation", School of Computer Science Xi'an Polytechnic University Xi'an China, 2009.
[5] Fang xu, "Research on a Predictive Model of the Customer Churn",Educational Institute, Lan Zhou University, Gansu Lanzhou, 730020,China, 2011.
[6] Wenbin Zhou, "Outlier Detection on Large-Scale Collective Behaviors", College of Computer Science and Technology Fudan University, Shanghai 200433, China, 2011.
[7] Chen Yu-bao, Li Bao-sheng, GeXin-quan, "Study on Predictive Model of Customer Churn of Mobile Telecommunication Company", School of Economics and Management, Beijing Information Science& Technology University, Beijing, 102613 China,, 2011.
[8] WitoldGruszczy□ski, "Application of Social Network to Improve Effectiveness of Classifiers in ChurnModeling", Institute of Control and Computations Engineering Warsaw University of Technology Warsaw, Poland, 2011.
[9]Lei Li, Fang-Cheng Shen, De-Zhang Yang, "Application of Data Mining Based on Rough Sets in the Field ofTelecommunications", College of Science, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China, 2011.
[10] ZHANG Yihua, "Vip Customer Segmentation Based on Data Mining in Mobile-communications Industry", Department of Information management, Business Administration college, Jimei UniversityXiamen, Fujian, P.R.China, 2010.