Volume 4, Issue 9, September 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Data Mining Techniques and Applications in Telecommunication Industry

Ms Ranju Marwaha
Assistant Professor/It
Ssiet Derabassi, Punjab, India

Abstract: Telecommunication companies routinely generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a rapidly changing and highly competitive environment. These companies also face a number of data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of their data, and the need to predict very rare events—such as customer fraud and network failures—in real-time

The main application areas of Business Intelligence and Data Mining in telecommunication industry include fraud detection, network fault isolation and improving market effectiveness.

Keywords: Data Mining, Telecommunications, Business Intelligence, Fraud Detection, Network fault Isolation, Marketing & CRM

I. INTRODUCTION

Data an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

A data mining algorithm is a set of heuristics and calculations that creates a data mining model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends. The algorithm uses the results of this analysis to define the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics.

The mining model that an algorithm creates from your data can take various forms, including:

- A set of clusters that describe how the cases in a dataset are related.
- A decision tree that predicts an outcome, and describes how different criteria affect that outcome.
- A mathematical model that forecasts sales.
- A set of rules that describe how products are grouped together in a transaction, and the probabilities that products are purchased together.

Microsoft SQL Server Analysis Services provides multiple algorithms for use in your data mining solutions. These algorithms are implementations of some of the most popular methodologies used in data mining. All of the Microsoft data mining algorithms can be customized and are fully programmable using the provided APIs, or by using the data mining components in SQL Server Integration Services.

You can also use third-party algorithms that comply with the OLE DB for Data Mining specification, or develop custom algorithms that can be registered as services and then used within the SQL Server Data Mining framework.

II. CHOOSING THE RIGHT ALGORITHM

Choosing the best algorithm to use for a specific analytical task can be a challenge. While you can use different algorithms to perform the same business task, each algorithm produces a different result, and some algorithms can produce more than one type of result. For example, you can use the Microsoft Decision Trees algorithm not only for prediction, but also as a way to reduce the number of columns in a dataset, because the decision tree can identify columns that do not affect the final mining model.

III. CHOOSING AN ALGORITHM BY TYPE

Analysis Services includes the following algorithm types:

- Classification algorithms predict one or more discrete variables, based on the other attributes in the dataset.
- Regression algorithms predict one or more continuous variables, such as profit or loss, based on other attributes in the dataset.
- Segmentation algorithms divide data into groups, or clusters, of items that have similar properties.
- Association algorithms find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis.
- Sequence analysis algorithms summarize frequent sequences or episodes in data, such as a Web path flow.

However, there is no reason that you should be limited to one algorithm in your solutions. Experienced analysts will sometimes use one algorithm to determine the most effective inputs (that is, variables), and then apply a different algorithm to predict a specific outcome based on that data. SQL Server data mining lets you build multiple models on a single mining structure, so within a single data mining solution you might use a clustering algorithm, a decision trees model, and a naïve Bayes model to get different views on your data. You might also use multiple algorithms within a single solution to perform separate tasks: for example, you could use regression to obtain financial forecasts, and use a neural network algorithm to perform an analysis of factors that influence sales.

IV. TYPES OF TELECOMMUNICATION DATA

Useful applications cannot be developed without understanding the various data used in telecommunication industries. So the first step in the Data Mining process is to understand the data. The different kinds of data used in this
industry are mainly grouped into 3 different types.

Call detail data

This is the information about the call, which stores as the call detail record. The number of call detail records generated is huge since every call is placed on the network, the details are stored. Call detail record includes information like originating and terminating phone numbers, date, time and duration of call. Usually these call detail records are not directly used for Data Mining
A list of features can be generated from the call detail data such as
Average call duration
Average number of call originated per day
Average number of call received per day
Percentage of no - answer calls
Percentage of day time calls (office hours)
Percentage of weekday calls
(Monday – Friday)

Network data

Telecommunication networks contain thousands of components, which are interconnected. These components are capable of generating error and status messages which leads to a large volume of network data. These network data are used for network management functions like fault detection. Expert systems have been developed to analysis these messages automatically, since the huge volume of network messages generated cannot be handles by technicians. Hence Data Mining technologies are used in identification of network faults by automatically extracting knowledge from network data. Network data is also generated in real time which can be accomplished by applying a time window to the data.

Customer data

Like any other business, telecommunication companies also have millions customers. Hence it is very much essential to have a database for storing the information about these customers. Information about the customer will include:

- ✓ Name of the customer
- ✓ Address details
- ✓ Payment history
- ✓ Service plan and so on

Group customer data is used to provide call detail data in order to identify fraud.

V. DATA MINING AND BI APPLICATIONS

The two main factors on which Data Mining and BI applications relay on include the availability of the problem that has to be approached and solved by the Data Mining and BI technologies and the availability of Data for implementing the technologies.

The main reason behind the significance of Data Mining and Business Intelligence
Applications in the Telecommunications industry are the availability of tremendously large volume of data.

Marketing and customer relationship management (CRM)

Telecommunication companies maintain a huge volume of data about their customers and their call details. This information can be used to profile the customers and these profiles can be used for marketing and forecasting purposes. The emphasis of marketing application in telecommunication industry has moved from identifying new customers to measuring customer value and then taking steps to return the profitable customers. This shift has happened because it is expensive to acquire new customers than retaining the existing ones. A numerous Data Mining methods can be used to generate the customer life time value (the total net income a company can expect from a customer over time) for telecommunication customers. Different Data Mining techniques are used to model customer life time value for telecommunication customers. The key element of modeling the life time value for a telecommunication customer is to estimate how long he/she will remain with their current network. It will help the company to predict when a customer is likely leave and to take proactive steps to retain the customer. One of the serious issues that the telecommunications industries face is the customer churn. The process that a customer leaving a company is referred to as churn and churn analysis can be done through numerous systems and methods.

Network Fault Isolation & Prediction Telecommunication networks are comprised of highly complex configurations of hardware and software. Since the industry requires optimum network efficiency and reliability, most of the network elements have the capability of self – diagnosis and generating status and alarm messages. Expert systems were developed to handle alarms . Network fault isolation in the Telecommunication industry is a quiet tedious task because of the
Following reasons.
Huge volume of data
A single fault can generate different unrelated alarms.
Hence alarm correlation has an important role in predicting network faults.

A proactive rapid response is very much essential for maintaining the reliability of the network. Data mining techniques like classification, neural network and sequence analysis can be used for identifying network faults. The telecommunication Alarm Sequence Analysis (TASA) is a Data Mining tool which support fault identification by searching for recurrent patterns of algorithms

This information can be used to generate a rule based alarm correlation system, which can be used for identifying faults in real time. Genetic algorithm is another method to predict the telecommunication switch failures .Time weaver is a genetic algorithm which has the capability to operate directly on the raw network level time series data. This algorithm will identify patterns that will successfully predict the target event. Bayesian Belief Networks can also be used to identify the network faults Standard classification tools can be used to generate rules to predict future failures but it has several draw backs. Most importance drawback of this is that some information will be lost in reformulation process.

Fraud Detection

Fraud is very serious issue that the telecommunication industry faces since it leads to the loss of revenue by billions of dollars. As provided by Gosset & Hyland 1999, the telecommunication fraud can be defined as ―any activity by which telecommunication service is obtained without intention of paying.

Telecommunication fraud can be classified into two categories namely

- ✓ Subscription fraud
- ✓ Superimposition fraud

Subscription fraud occurs when a customer opens an account with the intention of never paying. Telecommunication companies consider Superimposition frauds are the most significant problems which occurs when a perpetrator gains illegal access to the account of a legitimate customer. Both subscriptions fraud and Superimposition fraud should be detected immediately and customer account should be deactivated.

Cellular cloning was a very serious issue in 1990's. This was eliminated with the Authentication methods. Deviation detection and Anomaly detection are the most common techniques used for detecting superimposed fraud. Combined use of customer signatures dynamic clustering and pattern recognition are some other methods which are recently applied in this area. Absolute analysis and differential analysis are considered as the two main sub categories of approaches for fraud detection. According to , the most often used techniques for fraud detection in telecommunication include statistical modeling, Bayesian rules, visualization methods, clustering, rule discovery, neural network, Markov models as well as combinations of
more than one method. Customer data can also be used for detecting fraud. For example price plan and credit rating information can be in cooperated into the fraud analysis .

Another common method for fraud detection is to create a profile of customer's calling behavior and compare activity against this behavior. This calling behavior can be generated by briefing the call detail records for a particular customer. Fraud can be identified immediately after it happens, only if the call details records are updated in real time. Fraud detection system works at the customer level, not at the individual call level. Fraud detection application involves predicting a relatively rare event where the class distributions involved is highly twisted.

VI. CONCLUSION

Data Mining and BI applications play a significant role in the telecommunication industry due to the availability of large volume of data and the rigorous competition in the sector.

The primary application areas include marketing and Customer Relationship Management, Fraud detection and Network Management. The recent developments in the Data Mining and BI fields and the implementation and enhancement of existing techniques and methods ensure the continuous growth and compatibility of telecommunication companies that make use of them.

REFERENCES

[1] [1]. Liebowitz, J. (1988). Expert System Applications to Telecommunications. New York, NY: John Wiley

[2] [2]. Pareek, D.: Business Intelligence for Telecommunications. Auerbach Publications, Taylor & Francis Group LLC.

[3] [3].Aggarwal, C. (Ed.). (2007). Data Streams: Models and Algorithms New York: Springer.

[4] [4]. Weiss, G. (2004). Mining with rarity: A unifying framework SIGKDD Explorations.

[5] [5].Freeman, E., & Melli, G. (2006).

[6] [6].Mozer, M., Wolniewicz, R., Grimes, D., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunication industry

[7] M. Berry and G. Linoff. Mastering Data Mining. John Wiley and Sons, New York, USA, 2000.

[8] Gary Cokins, Ken King, "Managing Customer Profitability and Economic Value in the Telecommunication Indutry", SAS Institute White paper.

[9] Hangxia Ma, Min Qin, Jianxia Wang. (2009), "Analysis of the Business Customer Churn Based on Decision Tree Method", The Ninth International Conference on Control and Automation, Guangzhou, China.

[10] MO Zan, ZHOA Shan, LI Li, LIU Ai-Jun, 2007, "A predictive Model of Churn in Telecommunications Base on Data Mining". , IEEE International Conference on Control and Automation", Guangzhou, China.

[11] Yossi Ritcher, Elad Yom-Tov, Noam Slonim, "Predicting Customer Churn in Mobile Networks through Analysis of Social Groups". SIAM.

[12] PAKDD 2006 Data Mining Competition, http://www3. ntu. edu. sg/SCE/pakdd2006/competition/overview. htm