IJCSMS International Journal of Computer Science & Management Studies, Vol. 13, Issue 01, January 2013
ISSN (Online): 2231 –5268
www.ijcsms.com

DATA MINING AND MOBILE COMMUNICATION: AN OVERVIEW

Poonam Chaudhary
Department of computer science, Chaudhary Devi lal University, Sirsa, Haryana, India
poonam2_ch@yahoo.co.in

Dr. Dilbagh Singh
Department of computer science, Chaudhary Devi lal University, Sirsa, Haryana, India

Abstract
Mobile communication companies generate a tremendous amount of data. These data include call detail data, which describes the calls that traverse the mobile communication networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the mobile communication customers. This paper describes various fault detection techniques in the mobile communication with data mining. And describes how data mining can be used to uncover useful information buried within these data sets. Several data mining applications are described and together they demonstrate that data mining can be used to identify mobile communication fraud, improve marketing effectiveness, and identify network faults.
Keywords*: Mobile communications; fraud detection; marketing; network fault isolation.*

1. Introduction

The mobile communications industry was one of the first to adopt data mining technology. This is most likely because mobile communication companies mainly 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. Mobile communication companies utilize data mining to improve their marketing efforts, identify fraud and better manage their mobile communication networks. However, 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 popularity of data mining in the mobile communications industry can be viewed as an extension of the use of expert systems in the mobile communications industry (Liebowitz, 1988). These systems were developed to address the complexity associated with maintaining a huge network infrastructure and the need to maximize network reliability while minimizing labor costs. The problem with these expert systems is that they are expensive to develop because it is both difficult and time consuming to elicit the requisite domain knowledge from experts. Data mining can be viewed as a means of automatically generating some of this knowledge directly from the data.

1.1 Background

The data mining applications for any industry depend on two factors: the data that are available and the business problems facing the industry. This provides background information about the data maintained by mobile communications companies. The challenges associated with mining mobile communication data are also described .
Mobile communication companies maintain data about the phone calls that traverse their networks in the form of call detail records, which contain descriptive information for each phone call. In 2001, AT&T long distance customers generated over 300 million call detail records per day (Cortes & Pregibon, 2001) and because call detail records are kept online for several months, this meant that billions of call detail records were readily available for data mining. Call detail data is useful for marketing and fraud detection applications.

Mobile communication companies also maintain extensive customer information, such as billing information, as well as information obtained from outside parties, such as credit score information. This information can be quite useful and often is combined with mobile communication-specific data to improve the results of data mining. For example, while call detail data can be used to identify suspicious calling patterns, a customer's credit score

is often incorporated into the analysis before determining the likelihood that fraud is actually taking place.

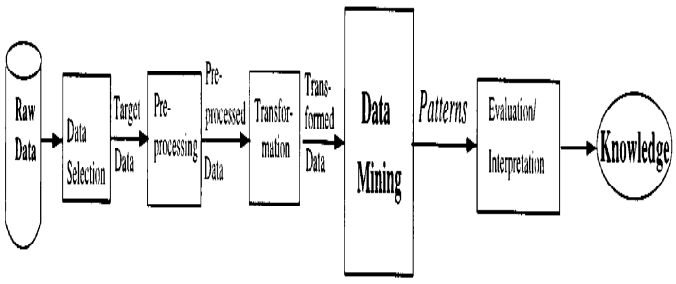
Mobile communications companies also generate and store an extensive amount of data related to the operation of their networks. This is because the network elements in these large mobile communication networks have some self-diagnostic capabilities that permit them to generate both status and alarm messages. These streams of messages can be mined in order to support network management functions, namely fault isolation and prediction. Thus, the scalability of data mining methods is a key concern. A second issue is that mobile communication data is often in the form of transactions events and is not at the proper semantic level for data mining. For example, one typically wants to mine call detail data at the customer (i.e., phone line) level but the raw data represents individual phone calls. Thus it is often necessary to aggregate data to the appropriate semantic level before mining the data. An alternative is to utilize a data mining method that can operate on the transactional data directly and extract sequential or temporal patterns. Another issue arises because much of the mobile communications data is generated in real-time and many mobile re-communication applications, such as fraud identification and network fault detection, need to operate in real-time. Because of its efforts to address this issue, the mobile communications industry has been a leader in the research area of mining data streams (Aggarwal,2007). One way to handle data streams is to maintain a signature of the data, which is a summary description of the data that can be updated quickly and incrementally.

A final issue with mobile communication data and the associated applications involves rarity. For example, both mobile communication fraud and network equipment failures are relatively rare. Predicting and identifying rare events has been shown to be quite difficult for many data mining algorithms (Weiss, 2004) and therefore this issue must be handled carefully in order to ensure reasonably good results.

1.2 Scope of data mining

Data mining derives its name from the similarities between searching for valuable business information in a large database, given databases of sufficient size and quality; data mining technology can generate new business opportunities by providing these capabilities.

1.3 Data mining process



1. Data cleaning- to remove noise and inconsistent data.
2. Data integration- where multiple data sources may be combined.
3. Data selection- where data relevant to the analysis
4. Data transformation- where data are transformed or consolidated into farms appropriate for mining by performing summary or aggregation operations.
5. Data mining – an essential process where intelligent methods are applied in order to extract data patterns
6. Pattern evaluation- to identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation- where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

1.4 Data mining life cycle

i. Find out the Problem – First of the entire problem is identified.
ii. Knowledge discovery- Given the problem, various reasons have to be analyzed by using one or more of the Data mining techniques. The exact solutions have to be finding out in order to resolve the problem, which we call it as the knowledge discovery.
iii. Implement the knowledge- based on above discovery, proper actions should be taken in or overcome the problem.

iv. Analyze the results- Once it has been implemented results need to be monitored and measured to find out outcomes of that action.

2. Related Work

Soman K.P. et al (2006) has described Data Mining for solving real-world problems (classroom learning). They explains how certain parameters of interest change as the algorithms proceeds. **U.P.Kulkarni** proposed a method that explores the capabilities of mobile agents to build an appropriate frame work and an algorithm that better suits the Distributed Data Mining applications. **Yang et al** developed and innovative sequential data mining system for mining the customers churning behaviors for the mobile communications industry. They have presented a novel model based clustering algorithm that integrates sequence clustering with multi dimensional data mining. Their method is a generalization of the 1st order Markov model-based clustering algorithm, which can deal with more natural data with additional attributes. .**Kappert, C.B.; Omta, S.W.F**. have described that Neural nets can prove to be viable option for data mining tasks. Their study has shown that the neural modeling task can be performed with in a time period comparable to traditional techniques. **Luo Bin Shao Peiji Liu Juan**, have described to overcome the limitations of lack of information of customers of Personal Handyphone System Service (PHSS) and to build an effective and accurate customer churn model, The empirical evaluation results suggest that customer churn models built have a good performance through the course of model optical selecting, and show that the methods and techniques proposed are effective and feasible under the condition that information of customers is very little and class distribution is skewed. **Louis Anthony Coxn** presented a new approach to demand forecasting that performs well compared to even sophisticated time series methods but that requires far less data. It is based on the following simple idea: Divide units of analysis (census blocks, customers, etc.) into groups with relatively homogeneous behaviors, forecast the behavior of each group (which can be done easily, by construction), and sum over all groups to obtain aggregate forecasts. Identifying groupings of customers to minimize forecast errors is a difficult combinatorial challenge that they address via the data-mining technique of classification tree analysis. **R,Sterritt and D W Bustard,** described progress toward automated fault identification through a fusion between these soft and hard computing approaches. Global telecommunication systems are at the heart of the Internet revolution. To support Internet traffic they have built-in redundancy to ensure robustness and quality of service. This requires complex fault management. The traditional hard approach is to reduce the number of alarm events (symptoms) presented to the operating engineer through monitoring, filtering and masking. The goal of the soft approach is to automate the analysis fully so that the underlying fault is determined from the evidence available and presented to the engineer. **Steinder,Malgoizata and Sethi,Adarshpal S.** applied Bayesian reasoning techniques to perform fault localization in complex communication systems while using dynamic, ambiguous, uncertain, or incorrect information about the system structure and state.

Sterritt, Roy has described two key fault management approaches:

(i) Rule discovery to attempt to present fewer symptoms with greater diagnostic assistance for the more traditional rule based system approach.

(ii)The induction of Bayesian Belief Networks (BBNs) for a complete "intelligent" approach.. **Li, J and Manikopoulos ,C** described prototyped a hierarchical, multi-tier, multi-window, soft fault detection system, namely the Generalized Anomaly and Fault Threshold (GAFT) system, which uses statistical models and neural network based classifiers to detect anomalous network conditions. In installing and operating GAFT, while both normal and fault data may be available in a test network, only normal data may be routinely available in a production network, thus GAFT may be ill-trained for the unfamiliar network environment. \

They present in detail two approaches for adequately training the neural network classifier in the target network environment, namely the re-use and the grafted classifier methods. The reuse classifier method is better suited when the target network environment is fairly similar to the test network environment, while the grafted method can also be applied when the target network may be significantly different from the test network.

3. Main Focus

Numerous data mining applications have been deployed in the telecommunications industry. However, most applications fall into one of the following three categories: marketing, fraud detection, and network fault isolation and prediction.

3.1 Mobile communications marketing

Mobile communication companies maintain an enormous amount of information about their customers and due to an extremely competitive environment have great motivation for exploiting this information. For these reasons the mobile communications industry has been a leader in the use of data mining to identify customers, retain customers, and maximize the profit obtained from each customer. Perhaps the most famous use of data mining to acquire new mobile communications customers was MCI's Friends and Family program. This program, long since retired, began after marketing researchers identified many small but well connected sub graphs in the graphs of calling activity (Han, Altman, Kumar, Mannila & Pregibon, 2002). By offering reduced rates to customers in one's calling circle, this marketing strategy enabled the company to use their own customers as salesmen. This work can be considered an early us of social-network analysis and link mining (Getoor & Diehl, 2005). A more recent example uses the interactions between consumers to identify those customers likely to adopt new mobile communication services (Hill,Provost & Volinsky, 2006). A more traditional approach involves generating customer profiles (i.e., signatures) from call detail records and then mining these profiles for marketing purposes. Over the past few years, the emphasis of marketing applications in the mobile communications industry has shifted from identifying new customers to measuring customer value and then taking steps to retain the most profitable customers. This shift has occurred because it is much more expensive to acquire new mobile communication customers than retain existing ones. Thus it is useful to know the total lifetime value of a customer, which is the total net income a company can expect from that customer over time. A variety of data mining methods are being used to model customer lifetime value for mobile communication customers.

A key component of modeling a mobile communication customer's value is estimating how long they will remain with their current carrier. This problem is of interest in its own right since if a company can predict when a customer is likely to leave, it can take proactive steps to retain the customer.

3.2 Mobile communications fraud detection

Fraud is very serious problem for mobile communication companies, resulting in billions of dollars of lost revenue each year. Fraud can be divided into two categories: subscription fraud and superimposition fraud (Fawcett and Provost, 2002). Subscription fraud occurs when a customer opens an account with the intention of never paying the account and superimposition fraud occurs when a perpetrator gains illicit access to the account of a legitimate customer. In this the fraudulent behavior will often occur in parallel with legitimate customer behavior (i.e., is superimposed on it). Superimposition fraud has been a much more significant problem for telecommunication companies than subscription fraud. Ideally, both subscription fraud and superimposition fraud should be detected immediately and the associated customer account deactivated or suspended. However, because it is often difficult to distinguish between legitimate and illicit use with limited data, it is not always feasible to detect fraud as soon as it begins. This problem is compounded by the fact that there are substantial costs associated with investigating fraud, as well as costs if usage is mistakenly classified as fraudulent (e.g., an annoyed customer). The most common technique for identifying superimposition fraud is to compare the customer's current calling behavior with a profile of his past usage, using deviation detection and anomaly detection techniques. The profile must be able to be quickly updated because of the volume of call detail records and the need to identify fraud in a timely manner. Cortes and Pregibon (2001) generated a signature from a data stream of call-detail records to concisely describe the calling behavior of customers and then they used anomaly detection to "measure the unusualness of a new call relative to a particular account." Because new behavior does not necessarily imply fraud, this basic approach was augmented by comparing the new calling behavior to profiles of generic fraud—and fraud is only signaled if the behavior matches one of these profiles. Customer level data can also aid in identifying fraud.

For example, price plan and credit rating information can be incorporated into the fraud analysis(Rosset,

Murad, Neumann, Idan, & Pinkas, 1999). More recent work using signatures has employed dynamic clustering as well as deviation detection to detect fraud (Alves et al., 2006). In this work, each signature was placed within a cluster and a change in cluster membership was viewed as a potential indicator of fraud.

There are some methods for identifying fraud that do not involve comparing new behavior against a profile of old behavior. Perpetrators of fraud rarely work alone. For example, perpetrators of fraud often act as brokers and sell illicit service to others—and the illegal buyers will often use different accounts to call the same phone number again and again. Cortes and Pregibon (2001) exploited this behavior by recognizing that certain phone numbers are repeatedly called from compromised accounts and that calls to these numbers are a strong indicator that the current account may be compromised.

A final method for detecting fraud exploits human pattern recognition skills. Cox, Eick & Wills (1997) built a suite of tools for visualizing data that was tailored to show calling activity in such a way that unusual patterns are easily detected by users. These tools were then used to identify international calling fraud.

Fraud applications have some characteristics that require modifications to standard data mining techniques. For example, the performance of a fraud detection system should be computed at the customer level, not at the individual call level. So, if a customer account generates 20 fraud alerts, this should count, when computing the accuracy of this system, as only one alert; otherwise the system may appear to perform better than it actually does (Rosset, Murad, Neumann, Idan & Pinkas, 1999). More sophisticated cost based metrics can also be used to evaluate the system. This is important because misclassification costs for fraud are generally unequal and often highly skewed (Ezawa & Norton, 1995). For this reason, when building a classifier to identify fraud, one should ideally know the relative cost of letting a fraudulent call go through versus the cost of blocking a call from a legitimate customer.

Another issue is that since fraud is relatively rare—and the number of verified fraudulent calls is relatively low—the fraud application involves predicting a relatively rare event where the underlying class distribution is highly skewed. Data mining algorithms often have great difficulty dealing with highly skewed class distributions and predicting rare events. For example, if fraud makes up only 2% of all calls, many data mining systems will not generate any rules for finding fraud, since a default rule, which never predicts fraud, would be 98.8% accurate. To deal with this issue, the training data is often selected to increase the proportion of fraudulent cases.

3.3 Network fault isolation

Mobile communication networks are extremely complex configurations of hardware and software. Most of the network elements are capable of at least limited self-diagnosis, and these elements may collectively generate millions of status and alarm messages each month. In order to effectively manage the network, alarms must be analyzed automatically in order to identify network faults in a timely manner or before they occur and degrade network performance. A proactive response is essential to maintaining the reliability of the network. Because of the volume of the data, and because a single fault may cause many different, seemingly unrelated, alarms to be generated, the task of network fault isolation is quite difficult. Data mining has a role to play in generating rules for identifying faults. The mobile communication Alarm Sequence Analyzer (TASA) is one tool that helps with the knowledge acquisition task for alarm correlation (Klemettinen, Mannila & Toivonen, 1999). This tool automatically discovers recurrent patterns of alarms within the network data along with their statistical properties, using a specialized data mining algorithm. Network specialists then use this information to construct a rule-based alarm correlation system, which can then be used in real-time to identify faults. TASA is capable of finding episodic rules that depend on temporal relationships between the alarms. For example, it may discover the following rule: “if alarms of type link alarm and link failure occur within 5 seconds, then an alarm of type high fault rate occurs within 60 seconds with probability 0.7.” Before standard classification tasks can be applied to the problem of network fault isolation, the underlying time-series data must be represented as a set of classified examples. This summarization, or aggregation, process typically involves using a fixed time window and characterizing the behavior over this window. For example, if n unique alarms are possible, one could describe the behavior of a device over this time window using a scalar of length n. In this case each field in the scalar would contain a count of the number of times a specific alarm occurs.

One may then label the constructed example based on whether a fault occurs within some other time frame, for example, within the following 5 minutes. Thus, two time windows are required. Once this encoding is complete, standard classification tools can be used to generate "rules" to predict future failures. Such an encoding scheme was used to identify chronic circuit problems (Sasisekharan, Seshadri & Weiss, 1996). The problem of reformulating time-series network events so that conventional classification based data mining tools can be used to identify network faults has been studied. Weiss & Hirsh (1998) view this task as an event prediction problem while Fawcett & Provost (1999) view it as an activity monitoring problem.

Transforming the time-series data so that standard classification tools can be used has several drawbacks. The most significant one is that some information will be lost in the reformulation process. For example, using the scalar-based representation just mentioned, all sequence information is lost. Timeweaver (Weiss & Hirsh, 1998) is a genetic-algorithm based data mining system that is capable of operating directly on the raw network-level time series data (as well as otter time-series data), thereby making it unnecessary to re-represent the network level data.

Given a sequence of time stamped events and a target event T, Time weaver will identify patterns that successfully predict T. Time weaver essentially searches through the space of possible patterns, which includes sequence and temporal relationships, to find predictive patterns. The system is especially designed to perform well when the target event is rare, which is critical since most network failures are rare. In this, the target event is the failure of components in the 4ESS switching system.

Conclusion

The mobile communications industry has been one of the early adopters of data mining and has deployed numerous data mining applications. The primary applications relate to marketing, fraud detection, and network monitoring. Data mining in the mobile communications industry faces several challenges, due to the size of the data sets, the sequential and temporal nature of the data, and the real-time requirements of many of the applications. New methods have been developed and existing methods have been enhanced to respond to these challenges. The competitive and changing nature of the industry, combined with the fact that the industry generates enormous amounts of data, ensures that data mining will play an important role in the future of the mobile communications industry.

References

[1] Cortes, C., Pregibon, D. Signature-based methods for data streams. Data Mining and Knowledge Discovery 2001; 5(3):167-182. Cortes, C., Pregibon, D. Giga-mining. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining; 174-178, 1998 August 27-31; New York, NY: AAAI Press, 1998.

[2] Ezawa, K., Norton, S. Knowledge discovery in telecommunication services data usin Bayesian network models. Proceedings of the First International Conference on Knowledge Discovery and Data Mining; 1995 August 20-21. Montreal Canada. AAAI Press: Menlo Park, CA, 1995.

[3] Aggarwal, C. (Ed.). (2007). *Data Streams: Models and Algorithms*. New York: Springer. Alves, R., Ferreira, P., Belo, O., Lopes, J., Ribeiro, J., Cortesao, L., & Martins, F. (2006). Discovering telecom fraud situations through mining anomalous behavior patterns. *Proceedings of the ACM SIGKDD Workshop on Data Mining for Business Applications* (pp. 1-7). New York: ACM Press.

[4] Baritchi, A., Cook, D., & Holder, L. (2000). Discovering structural patterns in telecommunications data. *Proceedings of the Thirteenth Annual Florida AI Research Symposium* (pp. 82-85). Cortes, C., & Pregibon, D (1998). Giga-mining. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 174-178). New York, NY: AAAI Press. Cortes, C., & Pregibon, D. (2001). Signature-based methods for data streams. *Data Mining and Knowledge Discovery*, 5(3), 167-182.

[5] Cox, K., Eick, S., & Wills, G. (1997). Visual data mining: Recognizing telephone calling fraud. *Data Mining and Knowledge Discovery*, 1(2), 225-231.

[6] Devitt, A., Duffin, J., & Moloney, R. (2005). Topographical proximity for mining network alarm data. *Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data* (pp. 179-184). New York: ACM Press.

[7] Fawcett, T., & Provost, F. (2002). Fraud Detection. In W. Klosgen & J. Zytkow (Eds.), *Handbook of Data Mining and Knowledge Discovery* (pp. 726-731). New York: Oxford University Press.

[8] Freeman, E., & Melli, G. (2006). Championing of an LTV model at LTC. *SIGKDD Explorations*, 8(1),27-32.

[9] Getoor, L., & Diehl, C.P. (2005). Link mining: A survey. *SIGKDD Explorations*, 7(2), 3-12. Hill, S., Provost, F., & Volinsky, C. (2006). Networkbased marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2), 256-276.

[10] Kaplan, H., Strauss, M., & Szegedy, M. (1999). Just the fax—differentiating voice and fax phone lines using call billing data. *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp.935-936). Philadelphia, PA: Society for Industrial and Applied Mathematics.

[11] Klemettinen, M., Mannila, H., & Toivonen, H. (1999).Rule discovery in telecommunication alarm data.*Journal of Network and Systems Management*, 7(4), 395-423.

[12] Sasisekharan, R., Seshadri, V., Weiss, S. Data mining and forecasting in large-scale telecommunication networks. IEEE Expert 1996; 11(1):37-43.

[13] Weiss, G. M., Hirsh, H. Learning to predict rare events in event sequences. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. 359-363.AAAI Press, 1998.

[14] Weiss, G. M., Provost, F. Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research 2003; 19:315- 354.

[15] Weiss, G. M., Ros, J, Singhal, A. ANSWER: Network monitoring using object-oriented rule. Proceedings of the Tenth Conference on Innovative Applications of Artificial Intelligence; 1087-1093. AAAI Press, Menlo Park, CA, 1998.