Application of Data Mining In Mobile Communications

Nisha Sharma[1] Dr. Chander Kant[2]

[1]Research scholar, Singhania University, Pacheri Bari, Jhunjhunu (Raj.), India

[2]Assistant Professor, Department of Computer Science and Applications, Kurukshetra University, Kurukshetra (Haryana)

Abstract – Mobile communication companies generate a tremendous amount of data. These data include call detail data, which describes the calls that pass through the networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the customers. This paper describes that how data mining can be used to uncover useful information covered within these data sets. Several data mining applications are described and together they demonstrate that data mining can be used to identify telecommunication fraud, improve marketing effectiveness, and identify network faults.

Key words: Telecommunications, fraud detection, marketing, network fault isolation.

----------♦----------

INTRODUCTION

The telecommunications industry generates and stores a tremendous amount of data. These data include call detail data, which describes the calls that traverse the telecommunication networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the telecommunication customers. The amount of data is so great that manual analysis of the data is difficult, if not impossible. The need to handle such large volumes of data led to the development of knowledge-based expert systems. These automated systems performed important functions such as identifying fraudulent phone calls and identifying network faults. The problem with this approach is that it is time- consuming to obtain the knowledge from human experts (the "knowledge acquisition bottleneck") and, in many cases, the experts do not have the requisite knowledge. The advent of data mining technology promised solutions to these problems and for this reason the telecommunications industry was an early adopter of data mining technology.

Telecommunication data pose several interesting issues for data mining. The first concerns scale, since telecommunication databases may contain billions of records and are amongst the largest in the world. A second issue is that the raw data is often not suitable for data mining. For example, both call detail and network data are time-series data that represent individual events. Before this data can be effectively mined, useful "summary" features must be identified and then the data must be summarized using these features. Because many data mining applications in the telecommunications industry involve predicting very rare events, such as the failure of a network element or an instance of telephone fraud, rarity is another issue that must be dealt with. The fourth and final data mining issue concerns real-time performance: many data mining applications, such as fraud detection, require that any learned model/rules be applied in real-time. Each of these four issues are discussed throughout this chapter, within the context of real data mining applications.

TYPES OF MOBILE COMMUNICATION 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 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.

1. 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. At a minimum, each

call detail record will include the originating and terminating phone numbers, the date and time of the call and the duration of the call. Call detail records are generated in real- time and therefore will be available almost immediately for data mining. This can be contrasted with billing data, which is typically made available only once per month.

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. Thus, the call detail records associated with a customer must be summarized into a single record that describes the customer's calling behavior. The choice of summary variables (i.e., features) is critical in order to obtain a useful description of the customer.

2. Network Data Telecommunication networks are extremely complex configurations of equipment, comprised of thousands of interconnected components. Each network element 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.

Due to the enormous number of network messages generated, technicians cannot possibly handle every message. For this reason expert systems have been developed to automatically analyze these messages and take appropriate action, only involving a technician when a problem cannot be automatically resolved.

3. 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 in Section 3 do not focus on this source of data. However, customer data is often used in conjunction with other data in order to improve results.

DATA MINING APPLICATIONS

The telecommunications industry was an early adopter of data mining technology and therefore many data mining applications exist. Several typical applications are described in this section. These applications are divided into three application areas: fraud detection, marketing/customer profiling and network fault isolation.

1. Fraud Detection

Fraud is a serious problem for telecommunication companies, leading to billions of dollars in lost revenue each year. Fraud can be divided into two categories: subscription fraud and superimposition fraud. Subscription fraud occurs when a customer opens an account with the intention of never paying for the account charges. Superimposition fraud involves a legitimate account with some legitimate activity, but also includes some "superimposed" illegitimate activity by a person other than the account holder. Superimposition fraud poses a bigger problem for the telecommunications industry and for this reason we focus on applications for identifying this type of fraud. These applications should ideally operate in real-time using the call detail records and, once fraud is detected or suspected, should trigger some action. This action may be to immediately block the call and/or deactivate the account, or may involve opening an investigation, which will result in a call to the customer to verify the legitimacy of the account activity.

The most common method for identifying fraud is to build a profile of customer's calling behavior and compare recent activity against this behavior. Thus, this data mining application relies on deviation detection. The calling behavior is captured by summarizing the call detail records for a customer, as described earlier in this chapter. If the call detail summaries are updated in real-time, fraud can be identified soon after it occurs. Because new behavior does not necessarily imply fraud, one fraud-detection system augments this basic approach by comparing the new calling behavior to profiles of generic fraud—and only signals fraud if the behavior matches one of these profiles. Customer level data can also aid in identifying fraud.

2. Marketing/Customer Profiling

Telecommunication companies maintain a great deal of data about their customers. In addition to the general customer data that most businesses collect, telecommunication companies also store call detail records, which precisely describe the calling behavior of each customer. This information can be used to profile the customers and these profiles can then be used for marketing and/or forecasting purposes.

We begin with one of the most well-known and successful marketing campaigns in the telecommunications industry: MCI's Friends and Family promotion. This promotion was initially launched in the United States in 1991 and, although now retired, was responsible for significant growth in MCI's customer base. The promotion offered reduced calling fees when calls are placed to others in one's calling circle. This promotion purportedly originated when market researchers noticed small sub graphs in the call- graph of network activity—which suggested the possibility of adding entire calling circles rather than the costly approach of adding individual subscribers (Han, Altman, Kumar, Mannila & Pregibon, 2002). It is worth noting that MCI relied primarily on its customers to bring in members of their calling circle, even though MCI could have utilized its call detail data to generate a list of the people in each calling circle. The most likely reason for this is that MCI did not want to anger its customers by using highly personal information (calling history). This demonstrates that privacy concerns are an issue for data mining in the telecommunications industry, especially when call detail data is involved.

3. Network Fault Isolation

Telecommunication 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 Telecommunication 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.

CONCLUSION

This paper described how data mining is used in the mobile communications 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).

Data mining applications must always consider privacy issues. This is especially true in the telecommunications industry, since telecommunication companies maintain highly private information, such as whom each customer calls. Most telecommunication companies utilize this information conscientiously and consequently privacy concerns have thus far been minimized. A more significant issue in the telecommunications industry relates to specific legal restrictions on how data may be used.

REFERENCES

1. Cortes, C., Pregibon, D. Signature-based methods for data streams. Data Mining and Knowledge Discovery 2001; 5(3):167-182.

2. 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.

3. Ezawa, K., Norton, S. Knowledge discovery in telecommunication services data using 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.

4. Fawcett, T., Provost, F. Adaptive fraud detection. Data Mining and Knowledge Discovery 1997; 1(3):291-316.

5. Fawcett, T, Provost, F. Activity monitoring: Noticing interesting changes in behavior.

6. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery

and Data Mining; 53-62. San Diego. ACM Press: New York, NY, 1999.

7. Han, J., Altman, R. B., Kumar, V., Mannila, H., Pregibon, D. Emerging scientific applications in data mining. Communications of the ACM 2002; 45(8): 54-58.

8. Kaplan, H., Strauss, M., Szegedy, M. Just the fax—differentiating voice and fax phone lines using call billing data. Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms. 935-936. Baltimore, Maryland. Society for Industrial and Applied Mathematics: Philadelphia, PA, 1999.

9. Klemettinen, M., Mannila, H., Toivonen, H. Rule discovery in telecommunication alarm data.

10. Mani, D. R., Drew, J., Betz, A., Datta, P. Statistics and data mining techniques for lifetime value modeling. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 94-103. San Diego. ACM Press: New York, NY, 1999.

11. Roset, S., Murad, U., Neumann, E., Idan, Y., Pinkas, G. Discovery of fraud rules for telecommunications—challenges and solutions. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 409-413, San Diego CA. New York: ACM Press, 1999.

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.