24

Data Mining in Telecommunications and Control

Objectives:

- Data mining in telecommunication industry helps to understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service.
- A large class of data mining algorithms developed for this purpose includes CART, C4.5, neural networks, and Bayesian classifiers, among others. One of the assumptions made by these algorithms, which are carried over into data mining applications, is that of clean data.
- The ability to handle noise in this case is obviously critical to the successful application of data mining algorithms; the treatment of noise typically falls short of handling the complete problem of data error.
- The companies in the telecommunications industry face the *problem of churning.* Churning is the process of customer turnover. This is a major concern for the companies having many customers who can easily switch to other competitors.
- Data mining is one solution to do appropriate credit scoring and to combat churns in the telecom industry.
- Data mining may be used in churn analysis to perform two key tasks: Predict whether a particular customer will churn and when it will happen; Understand why particular customers churn.
- Decision support in telecommunications forms the rules that can be used as decision support rules (for the exchange operator) or directly to automate the operation of the exchange.
- In control systems the purpose is to present a real-time knowledge processing (RTKP) procedure based on conjunctive and disjunctive matrices and operators.
- The field of knowledge discovery in databases (KDD) has delivered a variety of techniques to discover patterns from vast amount of data, which helps in mining for complex data.

S. Sumathi and S.N. Sivanandam: *Data Mining in Telecommunications and Control*, Studies in Computational Intelligence (SCI) **29**, 615–627 (2006)
www.springerlink.com

Abstract. The data mining applications in telecommunications industry, and a learning system for decision support in telecommunications case study, knowledge processing in control systems, and aircraft control case study are discussed in this section. A few scenarios where data mining may improve telecommunication services are discussed.

The deregulation of the telecommunications industry in many countries and the development of new computer and communication technologies and the telecommunication market are rapidly expanding and highly competitive. This creates a great demand for data mining in order to help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service.

In order to determine the needs of the telecommunication industry with respect to the data mining, an extensive literature survey analysis was performed at Telkom. They expressed a need for mining the data stored in the Telkom data warehouse. Almost all areas of Telkom's business can benefit from data mining, but in particular marketing and sales department. A serious problem for Telkcom, and for most companies in the telecommunications industry, is the *problem of churning.* Churning is the process of customer turnover.

A case study for decision support in telecommunications has been described. History data describing the operation of a telephone exchange is analyzed by the system to reconstruct understandable event descriptions. This case study is taken from Gerstner Laboratory, Czech Technical University, Czech Republic.

Real-time knowledge-based or knowledge-processing systems are playing an increasingly important role in transportation, manufacturing, control, and robotic and aerospace systems. They are no longer limited to low-level control functions. Control, supervision, and monitoring of complex hierarchical systems in dynamic and sometimes unpredictable or hazardous environments are typical tasks of current man-made systems.

Current development in real-time artificial intelligence is driven by a need to make knowledge-based systems work in real-time and a need to integrate knowledge-based approaches to handle the complexities of problem-solving behavior in control systems. The purpose is to present a real-time knowledge processing (RTKP) procedure based on conjunctive and disjunctive matrices and operators.

The case study taken from Sylvain Letourneau, University of Ottawa, Canada, is discussed to explain the how data mining is used for maintenance of complex systems. The anticipated contributions of this study were related to two fundamental problems in the fi eld of knowledge discovery in databases: i) automatic preparation of the data prior to model development and ii) use of diverse sources of information.

24.1 Data Mining for the Telecommunication Industry

The telecommunication industry has quickly evolved from offering local and long-distance telephone services to providing many other comprehensive communication services, including voice, fax, pager, cellular phone, images, e-mail, computer, and Web data transmission, and data traffic. The integration of telecommunication, computing network, Internet, and numerous other means of communications and computing is also underway.

The following are a few scenarios where data mining may improve telecommunication services.

24.1.1 Multidimensional Analysis of Telecommunication Data

Telecommunication data are intrinsically with dimensions such as calling time, duration, location of caller, and type of call. The multidimensional analysis of such data can be used to identify and compare the data traffic, system work load, resource usage, user group behavior, profit, and so on. For example, an analyst in the industry may wish to regularly view charts regarding calling source, destination, volume, and time-of-day usage patterns. Therefore, it is often useful to consolidate telecommunication data into large data warehouse and routinely perform multidimensional analysis using OLAP and visualization tools.

24.1.2 Fraudulent Pattern Analysis and the Identification of Unusual Patterns

Fraudulent activity costs the telecommunication industry millions of dollars a year. It is important to identify potentially fraudulent users and their atypical usage patterns; detect attempts to gain fraudulent entry to customer accounts; and discover unusual patterns that may need special attention, such as busy-hour, frustrated call attempts, switch and route congestion patterns, and periodic calls from automatic dial-out equipment (like fax machine) that have been improperly programmed. Many of these types of patterns can be discovered by multidimensional analysis, cluster analysis, and outlier analysis.

24.1.3 Multidimensional Association and Sequential Pattern Analysis

The discovery of association and sequential patterns in multidimensional analysis can be used to promote telecommunication services. For example, suppose we would like to find usage patterns for a set of communication services by customer group, by month, and by time of day. The calling records may be grouped by customer in the following form:

(Customer_id, residence, office, time, date, service_1, service_2,...)

A sequential pattern like "If a customer in the Los Angeles area works in a city different from her residence, she is likely to first use long-distance service between two cities around 5 pm and then use a cellular phone for at least 30 minutes in the subsequent hour every weekday" can be further probed by drilling up and down in order to determine whether it holds for particular pairs of cities and particular groups of persons (e.g., engineers, doctors, etc.). This can help promote the sales of specific long-distance and cellular phone combinations, and improve the availability of particular services in the region.

24.1.4 Use of Visualization Tools in Telecommunication Data Analysis

Tools for OLAP visualization, linkage visualization, association visualization, clustering, and outlier visualization have been shown to be very useful for telecommunication data analysis.

24.2 Data Mining Focus Areas in Telecommunication

The experience with applications of interest to the telecommunications business was carried out at Bell Atlantic STC, NY, where we focus on machine learning algorithms. A large class of data mining algorithms has developed out of ideas investigated earlier by researchers and developers of machine learning algorithms. Notable examples include CART, C4.5, neural networks, and Bayesian classifiers, among others. One of the assumptions made by these algorithms, which are carried over into data mining applications, is that of clean data.

All of these algorithms, and other like them, do relax the assumption from its strictest terms. They do not assume perfectly clean data, but rather assume that the data might be noisy. While the ability to handle noise is obviously critical to the successful application of data mining algorithms, the treatment of noise typically falls short of handling the complete problem of data error.

24.2.1 Systematic Error

Systematic errors arise in many applications, and they may be due to any of the following:

- Errors of calibration of instruments.
- Personal errors. These are errors caused by habits of individual observes.
- Imperfect technique.

We have found many examples of these in some of the telecommunications applications, investigated at Bell Atlantic Science and Technology Center, NY. One of these applications is classification of customer-reported telephone problems in the local loop of the telephone network. Problem diagnoses are high level, describing roughly that segment of the local loop where the trouble might be found, so that an appropriate technician might be dispatched to repair the trouble. The diagnoses are dispatched to the customer's premise: dispatch to the cable, dispatch to the central office, hold for further testing. The data describing the troubles includes information about the type of switch to which the customer's line is connected and electrical readings such as voltages and resistance, among others. The data mining problem here is to consider a large database of past troubles and their resolutions, and to develop

rules for sending the appropriate technicians out to fix problems that have a certain profile. The electrical readings that are a large component of the data are obtained via an automated line testing system. The line testing must be calibrated regularly, but in practice this rarely occurs. As a result, the system becomes miscalibrated, and all readings reported for a set of lines on a given day might be off by a systematic amount. Furthermore, the system's baseline readings can differ from day to day.

This source of systematic error is known, but there are no mechanisms in place to handle so that it can be eliminated from the data. Given the heavy load handled by the company, it is not clear that careful calibration can become a high priority item. Thus we can expect that the problem will persist. People can also affect the data in a systematic way. In particular, one source of the diagnoses for troubles is the technicians who fix the problems. They report results using a complex coding system. If a technician has memorized the wrong code to represent the outcome of a repair, it will be wrong consistently. Again, we have a good sense of the source of the problem, but it is not clear that it can be controlled. Also, aside from maintaining a profile of each technician, it is not clear that there is a mechanism that could automatically correct for these errors.

There are a number of different scenarios that arise with respect to systematic data error.

(1) The systematic error is well understood. In these cases, the data can be "cleaned" and data mining algorithms can be applied to the clean data.
(2) The errors can be reconciled. There are applications in which data may be obtained from several sources. In these cases, it may be possible to retain data that are consistent over the sources. This has the effect of cleaning the data by making the assumption that the data might have errors but that the errors will not be consistent over the various sources. We found that with the local-loop diagnosis application, we were able to use a variety of data sources to reconcile diagnostic error (through we were not able to account for calibration error).
(3) The data cannot be cleaned. These are cases where the error exists, but cannot be removed from the data. It is important to note that in these cases, the sources of the error might, in fact, be quite well known, but that additional complications make it difficult to pull the error out of the data.

One obvious reaction to these situations is to throw up our hands and assume that the application of data mining techniques will provide no useful results. But this reaction is unreasonable.

(1) If the amount of systematic error is small, or if the right algorithm is applied, the impact of the error might be small relative to other gains of the data mining.

(2) Data mining techniques might be useful for helping to identify systematic error, making the process of cleaning one's data a possibility.
(3) There are applications for which only a small amount of mined information can go long way to benefiting a company. In these cases, it is not in our best interest as data miners to simply dismiss an application as being "too hard." In the application described above, an improvement of only 1% over the current dispatch procedure could save the company over $3,000,000 annually.

More work needs to be done on:

(1) Developing data mining algorithms for cleaning systematic error out of data.
(2) Analyzing the tools we have so that we can determine how they are actually affected by different types of error.

24.2.2 Data Mining in Churn Analysis

Data mining is the discovery of knowledge from data, and uses a variety of tools ranging from classical statistical methods to neural networks and other new techniques originating from machine learning and artificial intelligence. Recently, data mining has been used with substantial results in enabling and improving database marketing, process optimization, and detecting fraud.

Almost all areas of Telkom's business can benefit from data mining, but in particular marketing and sales department. A serious problem for Telkcom and for most companies in the telecommunications industry is the *problem of churning.*

Churning is the process of customer turnover. This is a major concern for the companies having many customers who can easily switch to other competitors. Competition will become fierce in the next years as new companies enter the South African market.

Given the increase in the customer choice, there will be increase in the churn rate. Anderson Consulting recently estimated customer churn level of 30% per year in the cellular phone markets. The cost of churn in the telecomm industry is large.

A 1995 report from the Digital Equipment Corporation estimated the cost of churning in the wireless communication to be around $400 per new subscriber. It is clear that spending money holding on to existing customer is more efficient than acquiring the new customer.

In South Africa a further churn-related problem has been identified. Telkom, as requested by the South African Government, should install many phones in previously disadvantaged communities and homes. Because the Client in these areas is not financially self-sufficient, the churn problem aggravated. This is also an example of the closely related to the credit-scoring

problem, i.e., the decision to accept a client in the face of the associated risk involved. Data mining is one solution to do appropriate credit scoring and to combating churns in the telecom industry.

Data Mining may be Used in Churn Analysis to Perform Two Key Tasks

Predict whether a particular customer will churn and when it will happen;
Understand why particular customers churn.

These prediction and understanding tasks represent the two most important aspects of data mining in use today. By predicting which customers are likely to churn, the company can reduce the rate of churn by offering customers new incentives to stay. By understanding why customers churn the company can also work on changing their service so as to satisfy these customers pro-actively. In addition the data mining tools so as to choose the best strategy in terms of cost and effort can quantify the chance of the customer churning after action is taken.

24.3 A Learning System for Decision Support in Telecommunications – Case Study

We present a system for decision support in telecommunications. History data describing the operation of a telephone exchange is analyzed by the system to reconstruct understandable event descriptions. The event descriptions are processed by an algorithm inducing rules describing regularities in the events. The rules can be used as decision support rules (for the exchange operator) or directly to automate the operation of the exchange. This case study is taken from Gerstner Laboratory, Czech Technical University, Czech Republic.

In spite of the explosion of information technologies based on written communication, the most common and most frequently used tool is the telephone. Up-to-date private branch exchanges (PBX) provide comfort in managing the telephone traffic, namely regarding calls coming into an enterprise from the outside world. Communication proceeds smoothly provided that the caller knows with whom she wants to communicate and the person is available. In the opposite case, there is a secretary, receptionist, operator, or colleague that can for instance help to find a substituting person.

The operator is a person with no direct product, but with strong impact on productivity of other people. Despite that, a wide range of companies cancels the post of the telephone operator. The reason is that it is not easy to find a person who is intelligent enough to be good operator and to be modest enough to be just an operator. This opens area for computers – the computer is paid for only once so numbers fix costs set in. Moreover, the machine can work nonstop and provide additional data suitable for analysis allowing for improvements of the telecommunication traffic. Currently there are several

domains where computers are used in the PBX area (neglecting the fact that PBX itself is a kind of computer):

- *Automated attendant* – a device that welcomes a caller in a unified manner and allows him usually to reach a person, or choose a person from a spoken list; in both cases the calling party is required to co-operate.
- *Voice mail* – a device allowing to leave a spoken message to an unavailable person and some rather sophisticated methods of delivering the messages are available.
- *Information service* – the machine substitutes a person in providing some basic information usually organized into an information tree; the calling party is required to co-operate.

The aim of the above listed tools is to satisfy a caller even if there is no human service available at the moment. But all such devices are designed in a static, simple manner – they act always the same way. The reason is simple – they do not consider who is calling nor what they usually want – as opposed to the human operator. Comparing a human operator/receptionist to a computer, we can imagine the following the improvements of the automated telephony:

1. Considering who is calling (by the identified calling party number) and what number was dialed by the caller, the system can learn to determine the person most probably desired by the caller; knowledge can be obtained either from previous cases (taking into account other data like daytime, explicit information – long absence of some of the company's employee, etc.) or by "observing" the way how the caller was handled by humans before; this could shorten the caller's way to get the information she needs.
2. The caller can be informed by a machine in a spoken language about the state of the call and suggested most likely alternatives; messages should be "context sensitive."

Naturally, the finite goal of computerized telephony is a fully "duplex" machine that can both speak and comprehend spoken language so that the feedback with the caller can proceed in a natural dialog. We present a methodology where the goal is to satisfy the goal 1.

The task was defined by a telecommunication company that installs PBX switchboards in various enterprises. The experiments are based on the PBX logging data coming from one of the enterprises. The methodology is reflected in a unified system with inductive (learning) capabilities to produce decision support rules based on the data describing the previous PBX switching traffic. The system can be naturally adapted to the condition of a specific company (by including a formally defined enterprise-related background knowledge) as well as in the case of a change in the PBX firmware (again via an inductive process).

The language of Prolog (a subset of the language of first-order logic) is employed as a unified formalism to represent the input data, the background

knowledge, the reasoning mechanism, and the output decision support rules. The reason for this is the structured nature of the data with important dependence between individual records, and the fact that sophisticated paradigms are available for learning in first order logic. These paradigms are known as *inductive logic programming (ILP)*. The fundamental goal of ILP is the induction of first-order logic theories from logic and background knowledge. In the recent years, two streams of ILP developed, called the *normal setting* (where roughly theories with a "predictive" nature are sought) and the *nonmonotonic setting* (where the theories have a "descriptive" character).

24.4 Knowledge Processing in Control Systems

Several large real-time applications are required to operate in environments that are not fully structured. The lack of information and uncertainty of the environment requires the use of problems-solving techniques. Elevator group control is one such application. There are many possible situations comprising the state of all elevators, existing calls in the building, completion of previously scheduled cars, and combining new hall calls with performance criteria. There are several possible corresponding schedules and as new hall calls appear, the scheduled cars must be revised frequently. Since entering all the possible responses (schedules) into the computer is unfeasible, automating the response construction process will be required. Factory scheduling is another such application.

Looney proposed a matrix procedure for real-time knowledge processing considering only production rules with one antecedent. His procedure however does not preserve its matrix form when several antecedents are present in a rule. Furthermore, when handling multiple antecedents, the procedure is difficult to analyze and to predict if it could meet deadlines due to the chaining scheme adopted. This is a critical issue in real-time situations.

Alternate schemes based on a network type of representation for production rules are the RETE (C.L.Forgy, AI, in 1982 and EUREK (M.Funabashi et al., in 1988) procedures. However, both procedures lack the predictability property, which is essential in real-time applications.

More recently, Paul et al. developed an approach, which integrates problem-solving methodology and architectural primitives to reduce the variance at methodology level and at problem-solving level. Using this approach they have shown that problem-solving and real-time task coexist within a readily analyzable framework.

The purpose is to present a real-time knowledge processing (RTKP) procedure based on conjunctive and disjunctive matrices and operators. The proposed procedure affords the setting up of the focus of attention mechanisms and guarantees its respond time. Those are important characteristics that real-time knowledge-based systems should have.

24.4.1 Preliminaries and General Definitions

A typical RTKP system acting as a direct digital control system is shown in Fig. 24.1. The RTKP module is connected to information sources and receivers. Sources may be sensors connected to a process, human users, or even computer programs in large integrated systems. Receivers can be either actuators, human users, or again computer programs. The main idea behind this scheme is that RTKP takes information from a system, processes this information with the knowledge stored in it and then outputs new information to the system. Outputs are the control decisions. Figure 24.2 shows an RTKP system performing supervisory control tasks. The knowledge-processing task is encapsulated within a server to guarantee temporal isolation between it and conventional real-time tasks.

Internally, RTKP is divided into four basic parts. First is a preprocessor module responsible for the transformation of input information into the internal representation model is used. This module is also responsible for any mathematical treatment (by making transformation of variables, for example), as well as to preprocess task-dependent knowledge. The postprocessor module translates the internal representation model into output information in a former as required by the process. Between those two modules, there are the inference engine and the knowledge base. The last is the internal knowledge repository, coded in a usable format. Information provided by the preprocessor module and knowledge base is processed by the inference engine to generate the desired outputs.

The RTKP is defined by:

- Internal representations of information provided by the preprocessor module and information to be converted by the postprocessor module,
- An internal representation of the knowledge base,
- A procedure for the inference engine.

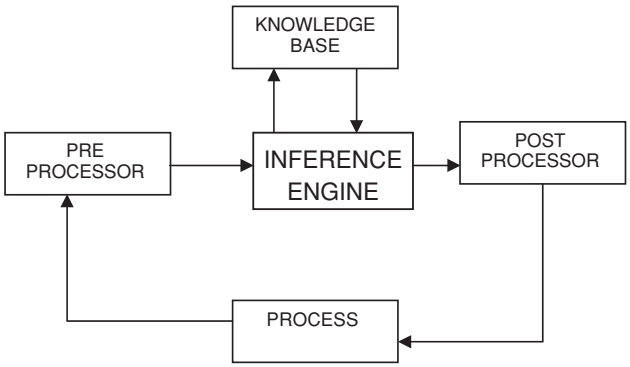

Fig. 24.1. Typical RTKP structure in direct digital control

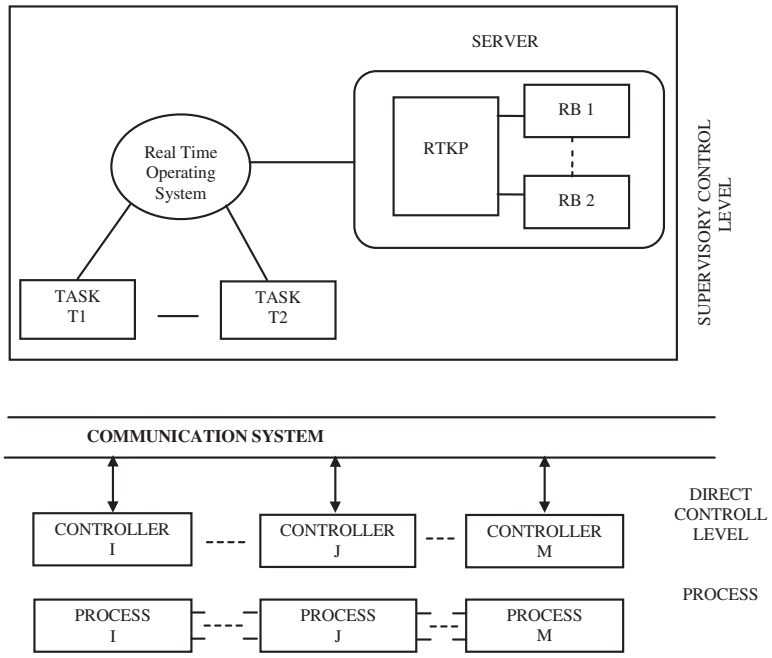

Fig. 24.2. RTKP in supervisory control systems

The knowledge base is composed of two parts: rule base and fact base. A fact is the truth value associated to particular proposition (or a term) used to store knowledge. A fact base is a set of terms, where each term has a meaning related to the process (an associated proposition). Propositions are considered within the prepositional logic framework, although they could also represent fuzzy propositions. In the proposed procedure, the fact base is represented by a fact vector where each component is related to term and contains its truth value.

The rule base can be viewed in two different representations. The first representation is for purposes of knowledge acquisition and analysis (called *the virtual representation*). It is presented as a collection of rules like: if <antecedent> then <consequent>, where <antecedent> represents a disjunctive association of terms. The second representation is coded version of the knowledge for processing purposes.

It is relevant now to review the main requirements that real-time knowledge processing systems should fulfill integration of numeric/symbolic processing, continuous operation, focus of attention mechanisms, interrupt manager services, optimum use of environment, response time warranty, temporal data processing, truth data maintenance, and the fact it dispenses explanatory modules.

24.5 Data Mining for Maintenance of Complex Systems – A Case Study

The operation and maintenance of modern sensor-equipped systems such as aircraft generates vast amounts of complex data. Proper use of this data to predict or explain component failures may lead to savings of several thousands of dollars, reducing the number of delays, and increasing the overall level of safety. The field of knowledge discovery in databases (KDD) has delivered a variety of techniques to discover patterns from vast amount of data. However, none of these techniques are designed to handle the diverse forms of data typically generated during the operation and maintenance of such complex systems. This case study is taken from Sylvain Letourneau, University of Ottawa, Canada. In this research, we study the specific issues to consider during the analysis of commercial aircraft data and process to handle these difficulties.

We aim at extracting useful information from large amounts of data collected from a fleet of 34 commercial aircraft over the last three years. Many issues with the analysis of this data have been identified (Letourneau et al., 1997). First of all, diverse sources and formats of data are to be considered. Available data includes: i) various types of sensor measurement reports describing the status of the aircraft in different phases of operation, ii) warning and failure messages generated when particular conditions occur, and iii) descriptions of aircraft problems along with the maintenance actions taken for each of them. Various sources of background knowledge are also available such as troubleshooting guides, training manuals, and empirical studies. A second difficulty comes from the complexity and the quality of the data. The number of parameters is high (i.e., often more than 100), several parameters are expected to have time-series relationships, and problems such as missing values, improper data types, and out-of-range data are frequently observed. Some sensor measurements must also be normalized due to the influence of variations in the environment.

We propose two major research directions to address these difficulties: i) development of a powerful data preprocessing approach to handle data complexity and data quality problems and ii) integration of the multiple sources of information described above to improve the quality of models learned from the data.

The data preprocessing approach proposed in the research should address the following tasks: i) normalization of the data according to the effects of contextual variations in the environment, ii) cleaning of the data, and iii) labeling of the instances so that supervised machine learning techniques can be used. We have developed the core of a novel domain-independent normalization technique that makes use of the analysis of variance (Letourneau.Matwin & Famili, 1998). Results from experiments with a large data set in the aircraft domain have shown that the proposed approach is powerful in reducing the

number of false alarms caused by random fluctuations in the environment. The current focus is on cleaning of the data and labeling of the instances.

The second major aspect of this work will address the use of different sources of information available during the operation and maintenance of complex systems. We studied the use of the domain information regarding the following aspects: i) extraction of the most appropriate features ii) analysis of meaningfulness of obtained models, and iii) improvement of the accuracy. This research is expected to enhance the process of knowledge discovery in databases so that it could be successfully applied to the maintenance and operation of complex systems.

24.6 Summary

This section has described data mining applications in telecommunications industry, and a learning system for decision support in telecommunications case study, knowledge processing in control systems and aircraft control case study.

Data mining can be applied fruitfully, as in network capacity utilization. In network capacity utilization, planning for telecommunication markets that expand, and the ability to service customers is highly affected by the capacity planning that has taken place long before. Data mining provides understanding of the underlying patterns and structures of service usage by customer groups. This insight allows capacity planners to optimize the investments in network facilities to better serve customers, while avoiding costly overexpansions, i.e., having enough capacity to deal with growing markets just at the right time.

24.7 Review Questions

1. How can data mining improve telecommunication services?
2. Write a short note on systematic error observed in mining.
3. How is data mining used in churn analysis?
4. Explain how data mining is used in PBX areas.
5. With typical structure explain real-time knowledge processing (RTKP) in direct digital control and supervisory control systems.