CHAPTER-28
Applications and Trends in Data Mining

28.Applications and Trends in Data Mining

"What are some specific examples of the use of data mining for applications in science and business? Where will data mining be in the future?" Here we discuss data mining applications and provide tips on what to consider when purchasing a data mining software system. Additional themes in data mining are described, such as visual and audio mining, statistical techniques for data mining, theoretical foundations of data mining, and intelligent query answering by the incorporation of data mining techniques. The social impacts of data mining and future trends are also discussed.

28.1 Data Mining Applications

In the previous chapters of this book, we have studied principles and methods for mining relational data, data warehouses, and complex types of data (including spatial data, multimedia data, time-series data, text data, and Web data). Since data mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and domain-specific, effective data mining tools for particular applications. In this section, we examine a few application domains and discuss how customized data mining tools should be developed for such applications.

Data Mining for Biomedical and DNA Data Analysis

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and advances in cancer therapies to the identification and study of the human genome by discovering large-scale sequencing patterns and gene functions. Since a great deal of biomedical research has focused on DNA data analysis, we study this application here. Recent research in DNA analysis has led to the discovery of genetic causes for many diseases and disabilities, as well as the discovery of new medicines and approaches for disease diagnosis, prevention, and treatment.

An important focus in genome research is the study of DNA sequences since such sequences form the foundation of the genetic codes of all living organisms. All DNA sequences are comprised of four basic building blocks (called nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T). These four nucleotides are combined to form long sequences or chains that resemble a twisted ladder.

Human beings have around 100,000 genes. A gene is usually comprised of hundreds of individual nucleotides arranged in a particular order. There are almost an unlimited number of ways that the nucleotides can be ordered and sequenced to form distinct genes. It is challenging to identify particular gene sequence patterns that play roles in various diseases. Since many interesting sequential pattern analysis and similarity search techniques have been developed in data mining, data mining has become a powerful tool and contributes substantially to DNA analysis in the following ways.

Semantic integration of heterogeneous, distributed genome databases:

Due to the highly distributed, uncontrolled generation and use of a wide variety of DNA data, the semantic integration of such heterogeneous and wide variety of distributed genome databases becomes an important task for systematic DNA coordinated analysis of DNA databases. This has promoted the

development of integrated data warehouses and distributed federated databases to store and manages the primary and derived genetic data. Data cleaning and data integration methods developed in data mining will help the integration of genetic data and the construction of data warehouses for genetic data analysis.

Similarity search and comparison among DNA sequences:

We have studied similarity search methods in time-series data mining. One of the most important search problems in genetic analysis is similarity search and comparison among DNA sequences. Gene sequences isolated from diseased and healthy tissues can be compared to identify critical differences between the two classes of genes. This can be done by first retrieving the gene sequences from the two tissue classes, and then finding and comparing the frequently in the diseased samples than in the healthy samples might indicate the genetic factors of the disease; on the other hand, those occurring only more frequently in the healthy samples might indicate mechanisms that protect the body from the disease. Notice that although genetic analysis requires similarity search, the technique needed here is quite different from that used for time-series data. For example, data transformation methods such as scaling, normalization, and window stitching, which are popularly used in the analysis of time-series data, is ineffective for genetic data since such data are nonnumeric data and the precise interconnections between different kinds of nucleotides play an important role in their function. On the other hand, the analysis of frequent sequential patterns is important in the analysis of similarity and dissimilarity in genetic sequences.

Association analysis: identification of co-occurring gene sequences:

Currently, many studies have focused on the comparison of one gene to another. However, most diseases are not triggered by a single gene but by a combination of genes acting together. Association analysis methods can be used to help determine the kinds of genes that are likely to co-occur in target samples. Such analysis would facilitate the discovery of groups of genes and the study of interactions and relationships between them.

Path analysis: linking genes to different stages of disease development:

While a group of genes may contribute to a disease process, different genes may become active at different stages of the disease. If the sequence of genetic activities across the different stages of disease development can be identified, it may be possible to develop pharmaceutical interventions that target the different stages separately, therefore achieving more effective treatment of the disease. Such path analysis is expected to play an important role in genetic studies.

Visualization tools and genetic data analysis:

Complex structures and sequencing patterns of genes are most effectively presented in graphs, trees, cuboids, and chains by various kinds of visualization tools. Such visually appealing structures and patterns facilitate pattern understanding, knowledge discovery, and interactive data exploration. Visualization therefore plays an important role in biomedical data mining.

Data Mining for Financial Data Analysis

Most banks and financial institutions offer a wide variety of banking services (such as checking, savings, and business and individual customer transactions), credit (such as business, mortgage, and automobile loans), and investment services (such as mutual funds). Some also offer insurance services and stock investment services.

Financial data collected in the banking and financial industries are often relatively completed, reliable, and of high quality, which facilitates systematic data analysis and data mining. Here we present a few typical cases.

Design and construction of data warehouses for multidimensional data analysis and data mining:

Like many other applications, data warehouses need to be constructed for banking and financial data. Multidimensional data analysis methods should be used to analyze the general properties of such data. For example, one may like to view the debt and revenue changes by month, by region, by sector, and by other factors, along with maximum, minimum, total, average, trend, and other statistical information. Data warehouses, data cubes, multifeature and discover-driven data cubes, characteristic and comparative analyses, and outlier analysis all play important roles in financial data analysis and mining.

Loan payment prediction and customer credit policy analysis:

Loan payment prediction and customer credit analyses are critical to the business of a bank, many factors can strongly or weakly influence loan payment performance and customer credit rating. Data mining methods, such as feature selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. For example, factors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus the total monthly income), payment-to-income ratio, customer income level, education level, residence region, credit history, and so on. Analysis of the customer payment history may find that, say, payment of-income ration is a dominant factor, while education level and debt ratio are not. The bank may then decide to adjust its loan-granting policy so as to grant loans to those whose application was previously denied but whose profile grows relatively low risks according to the critical factor analysis.

Classification and clustering of customers for targeted marketing:

Classification and clustering methods can be used for customer group identification and targeted marketing. For example, customers with similar behaviors regarding banking and loan payments may be grouped together by multidimensional clustering techniques. Effective clustering and collaborative filtering methods (i.e., the use of various techniques to filter out information, such as nearest neighbor classification, decision trees, and so on) can help identify customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing.

Detection of money laundering and other financial crimes:

To detect money laundering and other financial crimes, it is important to integrate information from multiple databases (like bank transaction databases, and federal or state crime history databases), as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of people, and so on. Useful tools include data visualization tools (to display transaction activities using graphs by time and by groups of people), linkage analysis tools (to identify links among different people and activities), classification tools (to filter unrelated attributes and rank the highly related ones), clustering tools, (to group different cases), outlier analysis tools (to detect unusual amounts of fund transfers or other activities), and sequential pattern analysis tools (to, characterize unusual access sequences). These tools may identify important relationships and patterns of activities and help investigators focus on suspicion cases for further detailed examination.

Data Mining for the Retail Industry

The retail industry is a major application area for data mining since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption and service records, and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability, and popularity of business conducted on the Web, or e-commerce. Today, many stores also have Web sites where customers can make purchases on-line, some businesses, such as Amazon.com,

exist solely on-line, without any bricks-and mortar (i.e., physical) store locations. Retail data provide a rich source for data mining.

Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.

A few examples of data mining in the retail industry are outlined as follows.

Design and construction of data warehouses based on the benefits of data mining: Since retail data cover a wide spectrum (including sales, customers, employees, goods transportation, consumption and services), there can be many ways to design a data warehouse. The levels of detail to be included may also vary substantially. Since a major usage of a data warehouse is to support effective data analysis and data mining, the outcome of preliminary data mining exercises can be used to help guide the design and development of data warehouse structures. This involves deciding which dimensions and levels to include and what preprocessing to perform in order to facilitate quality and efficient data mining.

Multidimensional analysis of sales, customers, products, time, and region: The retail industry requires timely information regarding customer needs, product sales, trends and fashions, as well as the quality, cost, profit, and service of commodities. It is therefore important to provide powerful multidimensional analysis and visualization tools, including the construction of sophisticated data cubes according to the needs of data analysis. The multifeature data cube, is a useful data structure in retail data analysis since it facilitates analysis on aggregates with sophisticated conditions.

Analysis of the effectiveness of sales campaigns: The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of dismounts and bonuses to promote products and attract customers. Careful analysis of the effectiveness of sales campaigns can help improve company profits. Multidimensional analysis can be used for this purpose by comparing the amount of sales and the number of transactions containing the sale items during the sales period versus-those containing the same items before or after the sales campaign. Moreover, association analysis may disclose which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.

Customer retention-analysis of customer loyalty: With customer loyalty card information, one can register sequences of purchases of particular customers. Customer loyalty and purchase trends can be analyzed in a systematic way. Goods purchased at different periods by the same customers can be grouped into sequences. Sequential pattern mining can then be used to investigate changes in customer consumption or loyalty, and suggest adjustments on the pricing and variety of goods in order to help retain customers and attract new customers.

Purchase recommendation and cross-reference of items: By mining associations from sales records, one may discover that a customer who buys a particular brand of perfume is likely to buy another set of items. Such information can be used to form purchase recommendations. Purchase recommendations can be advertised on the Web, in weekly flyers, or on sales receipts to help improve customer service, aid customers in selecting items, and increase sales. Similarly, information such as “hot items this week” or attractive deals can be displayed together with the associative information in order to promote sales.

28.2 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 other data traffic. The integration of telecommunication, computer network, Internet, and numerous other means of communication and computing is also underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is 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.

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

Multidimensional analysis of telecommunication data: Telecommunication data are intrinsically multidimensional with dimensions such as calling-time, duration, location of caller, and type of call. The multi-dimensional analysis of such data can be used to identify and compare the data; traffic, system workload, resource usage, user group behavior, profit, and so on. For example, analysis 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 warehouses and routinely perform multidimensional analysis using **OLAP** and visualization tools.

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 machines) that have been improperly programmed. Many of these types of patterns can be discovered by multidimensional analysis, cluster analysis, and outlier analysis.

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 you would like to find usage patterns for a set of communication services by customer group, by month, and by time of day. Customer in the following form may group the calling records:

A sequential pattern like "If a customer in the Los Angeles area works in a city different/row her residence, she is likely to first use long-distance service between two cities around 5 pm and then use a cellular phone/or 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 combination and improve the availability of particular services in the region.

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.

28.3 Review Questions

1 Expalin aboutData Mining Applications

2 Explain about Data Mining for the Telecommunication Industry

28.4 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition