Chapter 21
Management and Analysis Based on Data Mining

Han Zheng and Xiaobo Gao

Abstract As a kind of data analysis method and technology that finds out the potential information in a great deal of information, data mining has become the social focus. In the process of the information construction of the electric power industry, there is a great deal of historical data and it is urgent to apply the data mining technology to research and develop an analysis decision system to solve the key and prominent problems in the operation management of the power supply enterprises. This essay presents detailed comparison and analysis of the data mining algorithm. Based on the characteristics of the electric power management analysis, it focuses on discussing the clustering analysis algorithm. The electric power data management analysis system based on the data mining technology designed by this essay can process the data of mixed type and get good mining effect. The clustering analysis of the customer data of electric power can obtain good classifications and help to prediction customer's purchase behaviors.

Keywords Data mining · Electric power data · Management analysis · Clustering algorithm

21.1 Introduction

The fast development of the modern information technology leads an information wave globally. The channels to produce information are more and more. The information updates faster and faster. Hundreds of millions of data are produced in

H. Zheng (✉) · X. Gao
Department of Computer and Information Science, Hechi University, Yizhou 546300, China
e-mail: aqone@qq.com

Y. Yang and M. Ma (eds.), *Proceedings of the 2nd International Conference on Green Communications and Networks 2012 (GCN 2012): Volume 2*,
Lecture Notes in Electrical Engineering 224, DOI: 10.1007/978-3-642-35567-7_21,

various sectors [1]. However, the development and application of the data in the database is mainly search inquiry with inefficiency. Additionally, a considerable amount of data has very strong timeliness. The value of the data lowers rapidly with the time. Although simple data inquiry or statistics can meet some low level needs, what the people need is to find out the general knowledge that has guiding significance for various decisions from large quantity of data resources. This knowledge highly summarizes and abstracts a great deal of the data. But there is lack of means to discover the hidden knowledge in the data, which results in "data explosion but lack of knowledge" [2]. With the widely use of the database and computer networks and the using of the advanced automatic data generation and acquisition tools, the amount of the data owned by people has been growing sharply and mass data emerges in an endless stream [3]. For example, every day, up to 10,000 of customer purchase data is stored in POS system in super markets; every hour, various synchronous satellites send 50 siga (kilomega) bytes remote sensing image data to the earth. Obviously, a great deal of information can provide convenience the people, but at the same time, it brings about a series of problems. For example, too much amount of information is too much for people to master and digest; it is hard to distinguish whether some information is true or not, thus, it makes it difficult to correctly apply the information; different information organization forms result in that it is hard to together process the information effectively. These changes cause the traditional database technology and data processing means cannot satisfy the requirements. The rapid development of the Internet also makes various resources in the internet exceedingly rich, making it is like looking for a needle in a haystack to search information in the internet.

21.2 Data Mining Technology

Data mining has the branch of broad sense and narrow sense. From broad sense, data mining means the procedure of discovering the hidden, internal, and useful knowledge or information from a great deal of information. From narrow sense, data mining means a key step in knowledge discover an important step for taking useful model of establishing model.

The theory basis of data mining provides guidelines for developing and studying on it. As is known to all, the development and the exploitation of the data mining theory are related to many subjects. Data mining involves with machine learning, pattern recognition, statistics, intelligent database, knowledge acquisition, data visualization, high-performance calculation and expert system, and other fields. There have been many data mining products such as Business-object, SAS, Dbmines, and so on. Data mining is a kind of profound level method for analyzing the data. The idea platform for it (or called data inventory theory) is data warehouse (or data mart). People look the original data as the source to form the knowledge, like mining from the ore. The original data can be structured, such as the data in the database. It also can be half-structured, such as text, graph, image data, even the heterogeneous data distributed in the networks. The methods of discovering knowledge can be

mathematical and nonmathematical; can be syllogistic and inductive. The knowledge discovered can be used for information management, inquiry optimization, decision supporting and process controlling and so on. It also can be used for maintaining the data itself. Therefore, data mining is a general cross subject, which brings together researchers in various fields, especially the scholars and engineering technicians in the fields of database, artificial intelligence, mathematical statistics, visualization, and parallel computation and so on.

21.3 Data Mining Framework in Electric Power Data Management and Analysis

Data mining technique procedure is like mining or panning from the mine. Must determine where the gold mines for mining. Similarly, starting with the angle of practical applications, the whole data mining process must be based on the profound understanding the mining objects. Different objects require for using different data mining techniques. This essay combines the practical need in the electric power marketing system and establishes a data mining model as shown in Fig. 21.1.

21.4 Application and Realization of the Data Mining Algorithm in the Electric Power Data Management and Analysis System

Clustering analysis is a method to classify the data reasonably. It classifies the objects into groups or categories by certain rules. These categories are not given in advance but are determined based on the data characteristics. The target of

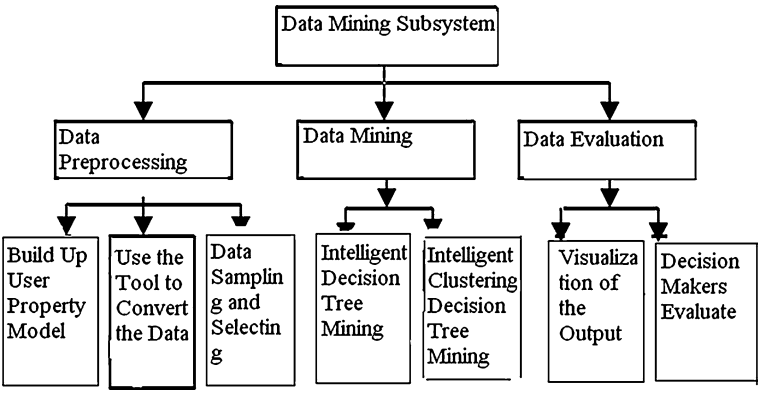

Fig. 21.1 Data mining model

clustering is to bring the data together into a category to minimize the similarity between the categories and maximize the similarity within category. The basic difference between classification problem (monitoring) and clustering problem lies in: in the classification problem, we know the classification property value of the training set, while in the clustering problem, we need to find out this classification property value in the training set. Clustering analysis is an important part of the multivariate statistical analysis. There have been multiple algorithms in the traditional statistical methods. With the emerging of the data mining technology, many algorithms have been put forward. At present, clustering analysis has been widely used in many fields, including pattern recognition, data analysis, image processing, and market study and so on.

Supposed that the data set is to be clustered includes n data objects. These data objects can be used to refer to persons, units, documents, countries and so on. Many inner-based clustering algorithms choose the two following typical data structures.

(1) Data matrix (or called the structure of the objects and the variables); it uses p variables (also called measure or property) to refer to n objects. For example, it uses age, height, weight, sex, and other properties to refer to the object "person". This kind of data structure is the form of the correlation chart, or can be looked as the matrix of $n * p$ (n objects * p variables).

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \tag{21.1}$$

(2) Dissimilarity matrix (or called object–object structure): store the similarity between every two objects among n objects, the expression form is one $n*n$ dimensional matrix.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2 & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix} \tag{21.2}$$

Here, $d(i, j)$ is the quantification expression of the dissimilarity between object i and object j. usually, it is a nonnegative value. The more similar or "closer" the object i and j are, closer to 0 is its value; more different the two objects are, larger

is its value. Due to $d(i, j) = d(j, i)$ and $d(i, i) = 0$, the simplified matrix above can be obtained.

The similarity means how similar the objects are. It is calculated based on the property value of the description objects. It is usually used to measure the similarity of the numerical variables. There are three common methods, as follows:

(1) Distance measurement: Murkowski distance is the standardized distance measure method in geometry problems.

$$\|x_a - x_b\| = \left(\left|x_{i1} - x_{j1}\right|^q + \left|x_{i2} - x_{j2}\right|^q + \ldots + \left|x_{if} - x_{jf}\right|^q\right)^{1/q} \tag{21.3}$$

When $q = 1$, it is called Manhattan distance; when $q = 2$, it is called Euclidean distance.

(2) Cosine measurement: use the angle between vectors or the cosine of the angle to measure the similarity.

$$s^{(c)}(x_a, x_b) = \frac{x_a^T x_b}{\|x_a\|_2 \|x_b\|_2} \tag{21.4}$$

One important characteristic of cosine measurement is not relying on the length of the vectors, namely when $a > 0$, it meets

$$s^{(c)}(ax_a, x_b) = s^{(c)}(x_a, x_b) \tag{21.5}$$

(3) Extending Jaccard coefficient similarity: duality Jaccard coefficient similarity is the ratio of the sharing property x_a AND x_b of the object and all the property x_a OR x_b they shall have. To extend it, obtain.

$$s^{(J)}(x_a, x_b) = \frac{x_a^T x_b}{\|x_a\|_2^2 + \|x_b\|_2^2 + x_2^T x_b} \tag{21.6}$$

21.5 System Design Analysis

This essay combines the practical situation of domestic electric power system and uses the advanced information technology to perform multi-layered, multi-angled and all-directional analysis and mining to design the electric power management and analysis decision supporting system. In this system, various data can be shown continuously, three-dimensionally and dynamically, which can be conveniently for

the managers to flexibly and fast extract and discover the useful information according to different demands to reveal the inner laws of the electric power markets and marketing, helpful for the managers to master at any time the electricity customer structure and customer characteristics. In this system, it mainly considers how to discover the information needed by decision according to the existing customer profiles. The problems of the electric power management and analysis that the data mining can solve include: market demand analysis and management, sale analysis of electricity power, important customer identification analysis, customer identification analysis, and comprehensive analysis of electric business and the analysis of other market behaviors.

Data filtering is that according to the content to be shown, choosing the smallest data set that meets the needs in which the noisy, polluted and incomplete data is to be eliminated. In this process, three aspects of the contents such as the time, angle, and original data are mainly considered. In order to reach certain purpose, people make an abstracted model for the prototype. In this system, the established model is that the original information is quantitatively and qualitatively analyzed and processed and then that it is converted into intuitive information to provide a basis for the decision makers to determine the best management decisions. Result shows the data in the established data model to the users to use by figurative ways. In which, the revolving pivot table which can complete getting and revolving functions shows the results quantitatively from the micro-aspect. While various statistic figures show the results from macro-aspect qualitatively. They also can complete the getting function.

In the process of realizing the electric power management and analysis system, lots of program codes are compiled, in which, the algorithm is realized through using JSP codes, partial codes are as follows:

```
//Defining local variables
Var zbz=Drop Down_zb. Item (Drop Down_zb. Selected Index);
Var per_a=Math.abs (TextBox_A.Text)/100;
Var per_b=Math.abs (TextBox_B.Text)/100;
Var per_c=Math.abs (TextBox_C.Text)/100;
Var per_d=Math.abs (TextBox_D.Text)/100;
//Setting figure facts
Active Document. Sections ["Customer Identification in Golden Period_ Total Sales Table"]. Facts. RemoveA11;
Active Document. Sections ["Customer Identification in Golden Period_ Total Sales Table"]. Facts. Add (zbz);
Active Document. Sections ["Customer Identification in Golden Period_ Area Figure"]. Facts. RemoveA11;
Active Document. Sections ["Customer Identification in Golden Period_ Area Figure"]. Facts. Add (zbz);
//Function
f_evaluate_theory_abcd (view, fact, result, order, per_a, per_b, per_c, per_d)
```

{Var i,j; Var exp; Var flag=1; Var pm_zyl=tem_ranking special col; Var pm_pml="ABCD rank col"; For (i=ActDoc. Sections [vie]. Columns. Count; >=1; i)…}

21.6 Conclusion

As a kind of data analysis method and technology that finds out the potential information in a great deal of information, data mining has become the social focus. In the process of the informatization construction of the electric power industry, there is a great deal of historical data and it is urgent to apply the data mining technology to research and develop an analysis decision system to solve the key and prominent problems in the operation management of the power supply enterprises. This essay presents detailed comparison and analysis of the data mining algorithm. Based on the characteristics of the electric power management analysis, it focuses on discussing the clustering analysis algorithm. The electric power data management analysis system based on the data mining technology designed by this essay can process the data of mixed type and get good mining effect. The clustering analysis of the customer data of electric power can obtain good classifications and help to prediction customer's purchase behaviors.

References

1. Chung H, Choi K, Chung H (2009) Generation of approximation rules using information gain. In: IEEE international fuzzy systems conference proceedings, vol 8(2). pp 22–25
2. Hong TP, Wang TT, Chien BC (2001) Learning approximate fuzzy ruels from training exples. IEEE international fuzzy systems conference, vol 8(9). pp 256–259
3. Pawlak Z (2006) Rough sets, rough relations and rough functions. Fundamenta informaticae 8(7):271–276