Application of Genetic Algorithm in Data Mining

Tan Jun-shan[1], He Wei[1], Qing Yan[2]
[1]College of Computer Science, Central South University of Forestry and Technology
[2] College of Materials Science and Engineering, Central South University of Forestry and Technology
Changsha, 410004, P.R. China
he163wei@163.com

Abstract—In this paper, the concept of data mining was summarized, and its significance that contributes to commerce was illustrated as well. Based on the characteristic, an important genetic algorithm which is widely used in data mining technology was proposed, whose principle and status were also expatiated. On the other hand, a best employee group of data mining algorithm based on genetic algorithm combining information management system for enterprise employees was proposed, the processes and steps which may solve problem used genetic algorithm as a example. Furthermore, the problems and challenges that data mining technology has to be faced with were discussed.

Keywords-genetic algorithm; data mining; search strategy; model

I. INTRODUCTION

Data mining is the non-trivial process that automatically collects those useful hidden information from the data muster, and is taken on as forms of rule, concept, rule and pattern and so on. It is useful for decision-makers owing to the virtues of analyzing historical and current data, discovering hidden relation and pattern, predicting possible behaviors which may occur in the future. The process of data mining is also called as process of knowledge discover, which is a new inter-disciplinary referring to such wide scope of subjects as databases, artificial intelligence, statistics, visualization, parallel computing and other fields [1]. From the technical point of view, data mining is the process which extracts information and knowledge which were latent among the rest and is hard to find out but is potentially useful from a large number of incomplete, noisy, ambiguous and actual data. From the commercial point of view, data mining is a novel business information process technology, its main feature is to conduct extraction, transformation, analysis and treatment in a large number of commercial business data, to extract key knowledge which can support the important business decision to be pregnancy, and to find out a similar business model from a these data. Since the emergence of ultra-large-scale data and advanced computer technology, practical needs of management and intensive computing power of these data urge data mining to produce and develop rapidly and be used widely. With the maturity of massive data collection technology, and development of more powerful computer processors and data mining algorithm, data mining technology attract more attention in commercial application [2]. Genetic algorithm is the most important technology in many mining technology, which can select information from large amounts of data in the data warehouse, find possible operating mode in market, and mining facts people hard to find out.

II. GENETIC ALGORITHM

Genetic algorithm is the random optimization method based on the principle of natural selection and biological evolution. Which changes the solution of problems into data individuals of a gene sting structure in genetic space by a certain coding scheme, converts objective function into fitness value, evaluates advantages and disadvantages of individuals, and as the basis to the genetic operation, It implements through the steps of identification of initial population, selection, cross, variation, evaluation and screening. Comparing with traditional optimization methods, the use of group search strategy, so information exchange between individuals of group and not dependent on the gradient information when research, processing characteristics of not dependent on problem model, suitable for parallel processing, with the strong ability of global search function solve problems, strong robustness etc. Now, It is used in mechanical engineering, electronic engineering, knowledge discovery, combinatorial optimization, machine learning, image processing, knowledge acquisition and data mining, adaptive control and artificial life, and other fields [3].

The major running steps of genetic algorithm are as follows:

1）Establish initial groups randomly with strings.

2）Calculate the fitness value of individuals.

3）According to genetic probability, to create new population by using the following operation.

a) Copy: Add existed excellent individuals copy to a new group, delete poor-quality individuals.

b) Hybrid: Exchange the two selected individual, the new individual of which will be added to the new group.

c) Variability: Random exchange a certain individual characters and then insert into a new group.

Repeat the implementation of hybrid and variability, choosing the best individual as the results of genetic algorithm once arrive to the conditions.

978-0-7695-3557-9/09 $25.00 © 2009 IEEE
DOI 10.1109/ETCS.2009.340

III. GENETIC ALGORITHM AND DATA MINING

A. Genetic algorithm in the position of data mining

Genetic algorithm plays an important role in data mining technology, which is decided by its own characteristics and advantages[4]. To sum up, mainly in the following aspects: 1) Genetic algorithm processing object not parameters itself, but the encoded individuals of parameters set, which directly operate to set, queue, matrices, charts, and other structure. 2) Possess better global overall search performance; reduce the risk of partial optimal solution. At the same time, genetic algorithm itself is also very easy to parallel. 3) In standard genetic algorithm, basically not use the knowledge of search space or other supporting information, but use fitness function to evaluate individuals, and do genetic Operation on the following basis. 4) Genetic algorithm doesn't adopt deterministic rules, but adopts the rules of probability changing to guide search direction.

B. The application of Genetic Algorithm in Data Mining

The following we will discuss the application that genetic algorithm finds the best employees from the database of information management for enterprise employee. The structure of information component data sheet for employee is shown in Tab.1.

TABLE I. STRUCTURE OF INFORMATION COMPONENT DATA SHEET

Employee ID	Age	Income level	Health condition	Gender	Frequency of comprehensive evaluation
199 801	20	Medium	General	M	8
199 802	34	Low	Good	F	5
199 803	24	Low	Good	M	6
199 804	28	High	Bad	F	4
199 805	43	Medium	General	M	9

Five chromosomes are used to define types of employees according to condition of employee data, namely: Minimum age of employee; Maximal age of employee; Income level, divided into high, medium and low, respectively expressed by 00, 01, 10; Health condition of employee, divided into good, general and bad, respectively expressed by 00, 01, 10; Employee gender, with 0 means women, 1 means man.

Individual chromosome coding, initial population, the fitness function, choice operator, cross-operator, mutation operator, and other key are needed to design in genetic algorithm; the main parameters are the group scale of n, cross-probability Pc, mutation probability Pm etc. These factors have great impact on the run performance algorithm, so must be designed carefully. Concrete algorithm and explanation as follows:

a) Coding strategy and coding string length L: Because of many parameters, multi-parameter coding technology can be used. Basic idea is to encode each parameter obtaining substring, and then combine these substrings into a complete chromosome. For example, 18 | 36 | medium | good | man gene strings express the employee group of age with 18 to 36 years old, medium-income, health condition is good, sex man, it will have a number of death genes if used binary code, so integer code and binary code are combined in use, such as 18 | 36 | medium | good | man encoded string of 18 | 36 | 01 | 00 | 1.

b) Select Operator: By using the selection mechanism of the certainty expected value model, that is expected value of integer part of $M = nf(x_i) / \sum_{1}^{n} f(x_i)$ to arrange the times that individual are selected, if selected to participate in cross-matching and, the survival expected value minus 0.5 in the next generation; Instead the survival expected value minus 1, then listing expected value of M of decimal part according the value from large to small , and one selection from large to small until the date is full. Such choice mechanism can overcome randomness in selection.

c) Cross-Operator: Because of multi-parameter coding technology is used, taking into the characteristics of string code, two cross is adopted.

*d) Mutation operator:*Adopting basic mutation operator, mutating age gene locus when below 5 random i*nteger.*

e) The group size M: When M for small value, which improves the evolution data of genetic algorithm, but decreases the diversity of group and might cause the premature phenomena of genetic algorithm; when M for greater value, which decreases the evolution speed of genetic algorithm. Therefore, comprehensive consideration in these two areas, the value of M for 20~100 is good.

f) Fitness function f(x): The best employee group, that is the employee group who obtains the highest number times in comprehensive evaluation in the same age condition, and the ultimate aim is to find young and excellent employee. In addition to adding a restrictive conditions: the minimum age of employee must be less than maximum age. The objective function can be set to (1).

$$g(x) = \exp(\frac{t(x)}{T}) + \exp(-\left|\frac{i(x)-5}{5}\right|) \qquad (1)$$

Thus, $t(x)$ accords with the times of comprehensive excellent evaluation of employee for x gene string; T is the total times of comprehensive excellent evaluation of all employee profits; $i(x)$ is age spacing of string of x.

Generally speaking, the choice intensity should be slight lower in the initial stage of genetic optimization, so as to avoid genetic groups have been controlled by one or a few individuals with higher fitness degree; in the latter of genetic optimization, because the difference is relatively small between groups, The potential ability is low if continue to optimize, it is necessary to improve choice intensity so as to constringe a better solution for genetic algorithm. So fitness function is designed to (2).

$$f(x) = 100\exp(-0.1g(x)) \qquad (2)$$

g) Cross-probability p_c : Cross-probability p_c control the frequency in exchange operation , high p_c can achieve greater solution space, thus reducing the stay in non-optimal solution on the probability, but large p_c will waste of much time in searching unnecessary solution space. To this end, the adaptive p_c can be used in (3).

$$P_c = \begin{cases} p_c * \dfrac{f_{\max} - x_{\max}}{f_{\max} - f_{avg}}, & \text{when } x_{\max} \geq f_{avg}, \ p_c \text{ as } 0.30 \\ p_c, & \text{when } x_{\max} < f_{avg} \end{cases} \quad (3)$$

Thus, x is the larger one in the operation of two individuals of cross-participation, $f_{\max}$ is the largest group fitness degree, f_{avg} is the average fitness degree.

h) Mutation probability p_m : Mutation probability p_m control of the new gene into the population ratio, if too low, some useful genes will not be able to enter the choice; if too high, too much random change, future generations may lose good characteristics inherited from both parents. To this end, the adaptive p_m can be used in (4).

$$P_m = \begin{cases} p_m * \dfrac{f_{\max} - y}{f_{\max} - f_{avg}}, & \text{when } y \geq f_{avg}, \ p_m \text{ as } 0.001 \\ p_m, & \text{when } y < f_{avg} \end{cases} \quad (4)$$

Thus, y is the individual fitness in a particular mutation operation, $f_{\max}$ is the largest group fitness degree, f_{avg} is the average group fitness degree.

i) Termination: When genetic algorithm runs to difference (| (f1-f2) / f1 | <ε) does not change or with small change between the two group generation of the best fitness degree, which is considered convergent and stop operation.

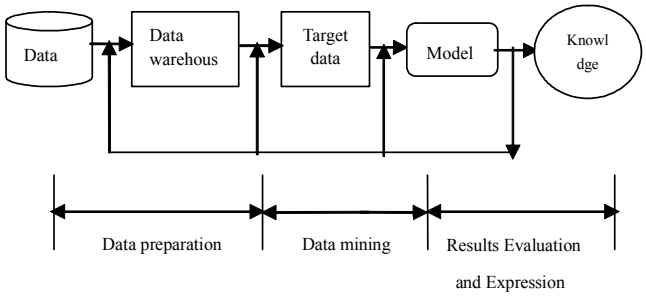

Figure 1. The process of data mining

In the process of data mining, data is usually needed to extract into data mining warehouse with data cleaning and pretreatment because of the complexity, and then begins to mine, the process of data mining is shown in Fig. 1.

Employee information is placed in a temporary table for pre-encoded through statements, and then extracts the data from the temporary table into mining warehouse.

IV. PROBLEM AND CHALLENGE IN DATA MINING TECHNOLOGY

Although data mining technology has a wide range of applications in various fields, its research is still not very mature, and there is great limitation in the application. The research of data mining technology need to continually develop in the future because of the above situation, some issue and challenge that data mining has to face in application process are as follows.

A. The efficiency of data mining

The more complicated relation among larger database, higher dimension and attribute, especially the development of large data warehouse and technology cause large amount data in the process of data mining. These factors will lead to high costs in searching knowledge. Currently, methods of parallel processing and sampling have been researched in order to obtain higher efficiency when dealing with large-scale data.

B. Data distortion

This issue is particular prominent in the field of business. At present, a novel method of the mining capacity noise by virtue of syncretizing statistics and fuzzy mathmatics to determine the hidden variables and their dependencies, so as to solve data mining issue of heterogeneous data.

C. Diversification of data

The current data mining tools is limited in the form of data processing, mainly used in the structure of the data. Text data mining, data mining, classification system, visualization system, space data system, distributed data mining and other new technology are the focus of the study and trends in current, however, the mining technology can not be well applied in practice[5]. In addition, the multimedia database has developed rapidly in recent years, and mining technology and software facing to multimedia database will become hot pot in research and development.

In the future, the data mining technology, with its the development, will be extensive and in-depth applied to human society in various fields, and will be development in the following areas: applicative exploration, scalable data mining, data mining and database system, data storage system and web database system integration, data mining language standardization, visualization of data mining, new mining method of complex data types, web mining, privacy and information security in data mining .

REFERENCES

[1] G. Q. Wang, D. Huang, "The summary of the date mining technology," Microcomputer Application Technology, vol. 2, 2007, pp. 9-14.

[2] J. Y. Liang, "The research of some common problems in the application of Genetic algorithms," Application Research of Computers, vol. 7, 1999, pp. 20-21.

[3] G. Y. Yu, Y. Z. Wang, "Applied Research of improved genetic algorithms," Machinery, vol. 5, 2007, pp. 58-60.

[4] D. L. Wang, M. Q. Li. "The application of data mining technology based on genetic algorithm," Journal of Nanchang University, vol. 1, A27, 2007, pp. 81-84.

[5] D. D. Li. "Data mining technology and development trend," Microcomputer Application Technology, vol. 2, 2007, pp. 38-40.