Information Systems and e-Business Management (2020) 18:747–765
https://doi.org/10.1007/s10257-018-00394-4

ORIGINAL ARTICLE

Data mining optimization model for financial management information system based on improved genetic algorithm

Wei Li[1] · **Qiling Zhou**[1] · **Junying Ren**[1] · **Samantha Spector**[2]

Received: 14 November 2018 / Revised: 23 December 2018 / Accepted: 26 December 2018 /
Published online: 14 January 2019

Abstract

The traditional corporate financial diagnosis method is susceptible to the choice of accounting policies, and there are serious lags, one-sidedness and limitations. A financial management information system based on improved genetic algorithm is proposed based on the financial management information system data mining and clustering analysis model framework, and based on the financial analysis related knowledge. By adopting the event-driven architecture, a financial management information system model based on data mining technology is constructed, which not only enables the data warehouse and data mining technology to play a role in decision support, but also enables the financial information and non-financial information of enterprises to be fully utilized. By extracting financial data, using the above decision tree classification algorithm for data mining, classifying tests according to subject categories and business processes, and evaluating the accuracy of the prediction results, and then determining whether the classification algorithm is selected. The test and analysis of the national tax financial analysis system were completed, and three public data sets and three national tax financial expenditure data sets were selected, and the algorithm was tested on the experimental platform. The test results show that the algorithm show good performance for large-scale data sets, especially financial expenditure data sets, and the test accuracy rate is not only stable but also maintains a relatively high range.

Keywords Genetic algorithm · Financial management system · Data mining · Clustering analysis · Model optimization

✉ Wei Li
13332267609@189.cn

[1] School of Maritime Economics and Management, Dalian Maritime University, Dalian, China

[2] School of Business, State University of New York at Albany, Albany, NY, USA

1 Introduction

With the development of information technology, a variety of application processing systems have emerged, resulting in a large amount of data information, which may contain extremely important knowledge. To discover the unknown and potential value hidden in the data, the traditional information processing methods obviously can not meet such high-level requirements. Therefore, it is necessary to make a deep analysis of the data in some way, extract the implicit information between the data, and form a set of analysis models, calculate and classify the data according to the quantitative indicators, and get the final analysis results (Chi et al. 2014; Tong et al. 2017; Zacharewicz et al. 2016). The process of data mining begins with defining specific problems, selecting mature algorithms according to the goal of data mining, extracting data from database or data warehouse, preparing and pre-processing data, and constructing data model. Evaluate the analysis data, adjust the structure of the model, reduce the errors of the actual system, and submit it to the relevant personnel for reference as assistant decision-making information (Singh 2017; Younus et al. 2015).

With the deepening of financial reform, China's banking industry is facing unprecedented challenges, banking reform is imperative. While carrying out the reform, the banks have also increased their investment in information system construction, in order to use the operation of these information systems, comprehensively integrate banking business and management data, improve the scientific accuracy of their decision-making, so as to enhance their core competitiveness (Xiao et al. 2014; Bider and Jalali 2016; Das and Padhy 2018). It brings many benefits, but there are also many problems, such as: the whole bank branches do not have a unified management system, various types of data do not have centralized management, the system of cost management is not perfect, the system is not open enough to communicate effectively with other system databases. With the intensification of market competition in recent years, various enterprises continue to improve the company structure, or attract investment, or implement the joint-stock system, the intensification of enterprise competition also makes the competition between commercial banks more intense (Engel et al. 2016; Iquebal et al. 2014). Commercial banks must strengthen all aspects of reform and innovation in order to adapt to the development and change of the market. Improving the financial management system of banks has become an inevitable choice for commercial banks.

Under the traditional architecture, the highly integrated financial information ultimately output by the financial information system can not usually meet the real-time information needs of users. At the same time, the demand for non-financial information is neglected. In this paper, using modern data warehouse and data mining technology, the author adopts business event-driven architecture, i.e. business events as the atomic unit of financial processing, to store the complete data of business events in the data warehouse, and to establish two subsystems at the same time (Macas 2014; Mikalef et al. 2017). Financial information processing subsystem can generate and provide various forms of financial and non-financial

information in real time according to the requirements of users. Financial management decision-making subsystem does not have to wait for the result data processed by the previous subsystem, but can directly access the data needed by the business process in the data warehouse (Guo et al. 2016; Cavuoti et al. 2014; Chang and Lin 2015). Genetic algorithm is a random search algorithm based on biological evolution theory. Because of its strong robustness, adaptability and implicit parallelism, it can quickly and effectively perform global optimization, which plays an important role in data mining technology. It has long been used in many aspects of artificial intelligence, and is one of the important methods for data mining and knowledge discovery. The algorithm is a global optimization algorithm, which is generally applied to find the optimal solution in a problem set, and is especially suitable for solving multi-objective optimization problems (Xie et al. 2015; Breuker et al. 2014; Zoet and Versendaal 2014). Drawing on the theory of biological evolution, the genetic algorithm simulates the problem to be solved as a process of biological evolution, using the fitness value to identify the individual's evolutionary ability, generating the next-generation solution by selecting operations such as crossover and mutation, and phasing out the fitness during the evolution process. A solution with a lower function value increases the solution with a higher fitness function value. After multiple recursive loops, it is very likely that the individual with the highest fitness function value will evolve after a certain stage of evolution.

There is a lot of information available in massive data. How can we use mature data mining technology and mining algorithm to obtain the relationship between financial data and the internal relationship of business from this information, form decision-making to provide leadership, help leadership deploy and carry out work? This will improve the efficiency and quality of the entire financial sector, which is the direction and purpose of this paper. Using advanced information analysis tools and deep data mining analysis methods, according to the direction of business data flow, the data information with decision-making nature in the national tax financial data is excavated (Kaiser et al. 2013). That is to say, from simple query to knowledge mining from financial data, the knowledge found has specific preconditions and constraints, all of which are specific to financial analysis of state tax. The financial data analysis system will extract data sets from the database of the national tax financial system, and form knowledge decision-making according to the application scope of the data of accounting subjects, thus separating leaders and cadres from the complicated business data and providing assistant decision-making management for leaders at all levels (Wu 2015; Wu et al. 2018). Timely budget and allocation for the next year, improve the utilization of state tax financial data, and standardize the implementation of the financial system to make great contributions.

2 Overview of data mining technology

2.1 Data mining

In the world of massive data information, data mining is the process of mining useful knowledge from these data. With the development of information technology, a large number of data and information accumulate, and find valuable, non-dominant information in a large number of databases, and then make decisions based on these information. It enables many organizations, such as banks, financial services institutions, hospitals, retail stores and government agencies, to obtain a large number of available data sets, databases or data warehouses. In order to obtain the maximum value of these databases or data warehouses, data mining is an essential part of making key business decisions (Ke et al. 2018; Ng and Khor 2015). Data mining is the process of extracting useful information and knowledge hidden in a large number of incomplete, noisy, fuzzy and random data, which people do not know beforehand, but have potential. Data sources are diverse, and must be real, massive and noisy. Knowledge is the result of data mining. It is based on data sources, analyzing and reasoning data, obtaining rules, rules, assertions and so on, and discovering knowledge that users are interested in. The specific flow chart is as shown in Fig. 1.

The knowledge discovered by data mining usually includes the following kinds: generalized knowledge, which is a general description of the common characteristics of the same kind of things. What is discovered is a general, higher level and macro knowledge, which is the abstraction, generalization and refinement of data. Differential knowledge reflects the difference in nature between different things.

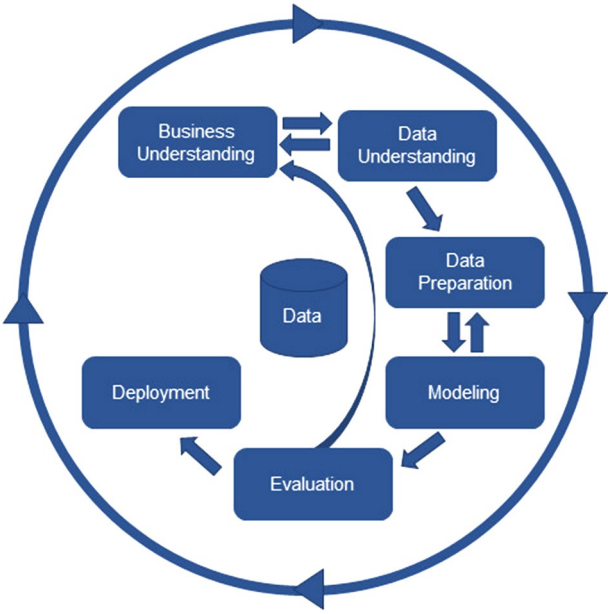

Fig. 1 Sketch of data mining process

Predictive knowledge predicts future data based on existing data. Relevant knowledge reflects the relationship between different things. If there is some relationship among multiple attributes, other attribute values can be predicted according to the attribute values of one item. Clustering knowledge is to classify data sets according to the common trends and patterns of data, so that the similarity of data within a class is the highest, and the difference of data between classes is the greatest.

The process of data mining can be divided into the following steps (Kridel and Dolk 2013):

1. Goal definition
   A clear and clear business goal is the key to the success of any data mining project. For some projects, its business may be relatively clear, but for most data mining projects, it is difficult to give a very clear business objectives. Therefore, as a business decision maker of a project, it is very important to make a clear description of the relevant business logic and mining objectives before the project starts. Otherwise, aimless data mining is clearly not successful.
2. Data processing
   Data processing mainly includes data collection and selection. Data collection methods can be varied, mainly depending on the open source of data and project funding considerations. Some data can be purchased directly from specialized data collection companies, such as census data, household registration data and personal file data. However, some data acquisition is more troublesome, if we need data acquisition difficulties, or even the relevant data does not exist at all, then it is imperative to investigate and collect data independently. The goal of data mining determines how to collect data. The main task of this step is to identify and collect data that can be used for mining.
3. Data conversion
   After the above two steps, in most cases, data is also needed to be transformed. The transformation method is decided mainly by data mining type, mining tool and mining technology. Several typical transformation methods: reorganization classification, symbol conversion into numerical value, mathematical transformation and so on. When data transformation is completed, we can do data mining.
4. Data mining
   Data mining is the core of this project. Generally speaking, the main data mining methods are clustering analysis, prediction model, time series analysis and link analysis.
5. Outcome assessment
   The result evaluation refers to transforming the results of data mining into business value, that is, based on the knowledge extracted by data mining, providing reasonable advice for the actual business and transforming it into a model that the user can understand.

The basic steps are shown in Fig. 2.

The task of data mining is to mine potential rules, patterns and knowledge from massive historical data. Its task determines its function. Generally speaking,

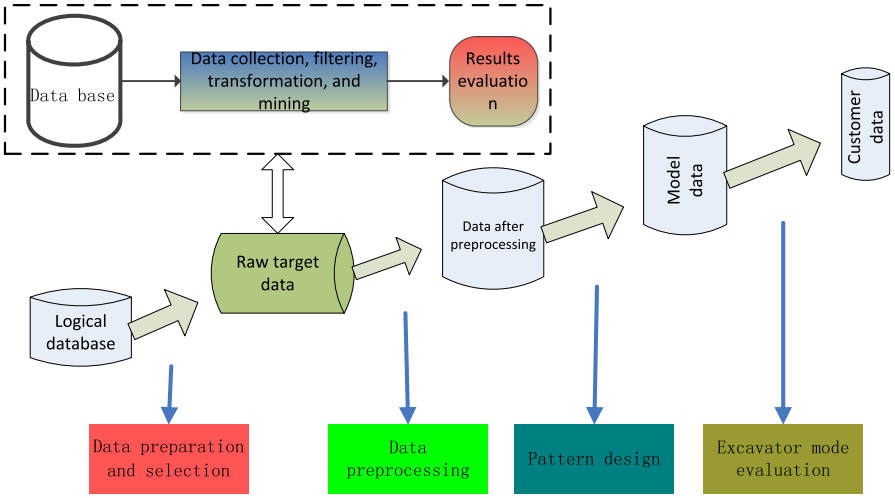

Fig. 2 Data mining process

data mining technology has two basic functions, that is, description function and prediction function. There are 4 main tasks of data mining in task Mining:

1. Concept description, simplifying and generalizes data characteristics, and generalizes the connotation of certain objects.
2. Find Association rules, analyze some correlation between objects or data, and get corresponding rules.
3. Clustering, i.e. clustering similar objects, has the greatest similarity among similar objects and the smallest similarity among different objects.
4. Deviation detection can find out the potential valuable difference between the results and reference values, so as to obtain useful knowledge.

After years of development and Research on data mining, more and more technologies are used in data mining system, and the accuracy of the results is getting higher and higher. Because, for a technology that is not quite adapted, other methods may work. This is mainly determined by the type and size of the data and the type of the problem.

It is believed that with the rapid development of computer technology and the increasing complexity of various businesses, the types of data are becoming more and more complex, and the amount of data is becoming larger and larger, and data mining will play an increasingly important role. Different types of data mining use many different technologies. Even different technologies may be required for the same type of data mining. What technology is used in data mining depends on business issues, data structures, and available computer software packages. The most commonly used data mining techniques include decision trees, association rules, cluster analysis, statistical analysis, and rough sets.

2.2 Genetic algorithm

The genetic algorithm originated from the study of adaptive behavior of natural and artificial systems in the 1960s. It was first proposed by John Holland, a professor at the University of Michigan in the United States. It is a simulation of biological genetic and evolutionary processes in the natural environment, an adaptive probabilistic search algorithm for global optimization form, we can solve the optimization problem and provide an effective way. The genetic algorithm uses a bit string encoding technology, the need for specific problems to generate the initial population, then the population fitness evaluation and selection, crossover and mutation, a series of genetic operation. The process of genetic operation based on adaptive method of value selection of excellent individuals in the proportion of the current population based, to produce the next generation population through crossover and mutation operation, until the desired conditions are thus inherited from generation to generation.

Genetic algorithm based on population way to organize search at the same time, so you can search the solution space in multiple areas, particularly suitable for massively parallel processing. The genetic algorithm is self-organizing, adaptive characteristics, using natural selection and simple genetic manipulation, it processes the object parameter encoding set instead of the parameters of the problem itself, so it can make the calculation process is not limited by the constraints of the search space, and does not need other auxiliary information. Therefore, genetic algorithm can not only achieve higher search efficiency, but also has the characteristics of simplicity and ease of operation.

The characteristics of the genetic algorithm has no requirements on state function, genetic algorithm for a specific problem only after a simple modification can be applied in other problems. If some domain knowledge and then adding specific problems, or with existing relevant algorithms, it can solve a complex problem, and genetic algorithm has good versatility and extensibility. At present, with the development of computer technology, genetic algorithm has got more and more attention, and in machine learning, pattern recognition, image processing, combinatorial optimization, VLSI design and optimization control has been successfully applied. Genetic algorithm is a random search algorithm. It can effectively use useful information to guide the search by evaluating the fitness of individuals and the role of genes in individuals.

The basic idea of genetic algorithm is as follows:

Step 1 Initialize the operation parameters;
Step 2 The initial population is generated randomly.
Step 3 Calculate the appropriate value of individual population.
Step 4 When the termination conditions are not met, two individuals are selected from the father generation, a series of genetic operations (crossover and variation) are performed, a new generation of population is generated, and a new generation of population is evaluated.

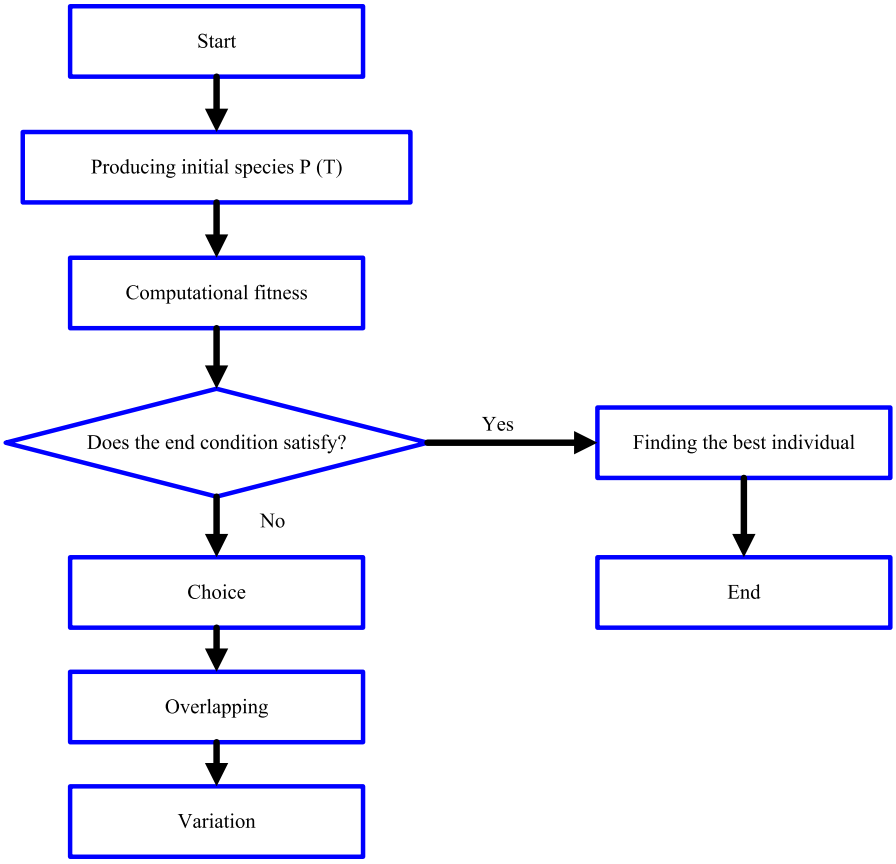

Fig. 3 Basic genetic algorithm structure

As shown in Fig. 3, genetic algorithm can provide a better general framework for solving complex system optimization problems, which does not depend on specific domain knowledge and problem types. Generally, the design of genetic algorithm has five basic components:

1. Determine the genetic expression of the solution of the problem, that is, choose the coding scheme of the problem.
2. The method of creating initial population of solution.
3. Determine the fitness function based on the appropriate value of individuals, which is the fitness function.
4. Determine the genetic operators of the next generation of individuals in the process of genetic manipulation.
5. The determination of the relevant operational parameters of the algorithm. The selection of appropriate functions and the design of genetic operators are the main

problems to be considered in the construction of genetic algorithm, and also the key steps in the specific design of genetic algorithm.

2.3 Cluster analysis

Clustering mining is a very important part of data mining, and it is also one of the main tasks of data mining. Clustering is to classify some objects. The similarity of objects belonging to the same class is as great as possible, and the differences between objects belonging to different classes are as great as possible. The results of cluster analysis can also provide some references for association mining in data mining. A cluster is a data record with similar characteristics or a subset of the customer. The process of gathering data records with similar characteristics and identifying them in this way, or grouping customers and data records, is called clustering. The goal of clustering is to divide data into group processes. We study how to classify objects into several classes without training.

In natural science and Social Sciences, there are a large number of classification problems. The so-called category refers to the set of similar elements. Clustering is a process of classifying data into different classes or clusters, so the objects in the same cluster are very similar, while the objects in different clusters are very different. Clustering analysis, also known as group analysis, is a statistical analysis method for classification problems. Clustering analysis is an analysis method in data mining. It can analyze data distribution independently and divide similar data objects into clusters according to the characteristics of data. Clustering analysis is applied to various fields, such as data mining, pattern recognition, statistics, machine learning and so on. The goal of cluster analysis is to classify data on a similar basis. Clustering technology is applied in various fields. It is mainly used to measure the similarity between different data and to divide data objects into different clusters. In business, market analysts use clustering technology to analyze consumers' consumption records, and conclude different types of consumers' consumption patterns, so as to realize the distinction between different consumer groups. In biology, clustering is used to analyze the classification of animals and plants, as well as clustering gene data to find genes with similar functions.

In clustering algorithm, the measure of the difference between samples is based on the distance in feature space rather than the similarity between samples. In order to quantify the similarity of samples conveniently, a measure of sample space distance can be metric or semi-metric. Dissimilarity is usually called distance, which can be represented by $d(x, y)$. When x and y are similar, the value of $d(x, y)$ is very small. When x and y are different, the value of $d(x, y)$ is very large. There are three commonly used distance functions:

1. Minkowski distance

   Suppose that n is the dimension of characteristic. x and y are corresponding characteristics. The Minkowski distance formula is:

$$d(x,y) = \left[\sum_{i=1}^{n} |x_i - y_i|^r\right]^{1/r} \tag{1}$$

When r takes different values, the Minkowski distance formula evolves into some special distance measurement formulas.

When r = 1, the formula (1) is the absolute distance. That is:

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

When r = 2, the formula (1) is the Euclidean distance. That is:

$$d(x,y) = \left[\sum_{i=1}^{n} |x_i - y_i|^2\right]^{1/2} \tag{3}$$

2. Two order distance

Suppose that x and y are corresponding characteristics. The formula for the two degree distance measure of x and y is as follows:

$$d(x,y) = ((x-y)^T A(x-y))^{1/2} \tag{4}$$

Among them, A is a nonnegative definite matrix.

Similarly, the two type distance metric formula can evolve into some special distance formulas.

When A is E of the unit matrix, the two type distance formula evolves into Euclidean distance formula.

$$d(x,y) = \left[\sum_{i=1}^{n} |x_i - y_i|^2\right]^{1/2} \tag{5}$$

When A is diagonal matrix, the two type distance formula evolves into the weighted Euclidean distance formula.

$$d(x,y) = \left[\sum_{i=1}^{n} a_{ii}|x_i - y_i|^2\right]^{1/2} \tag{6}$$

3. Cosine distance

The formula for cosine distance is as follows:

$$d(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{7}$$

The distances between the two classes have the following measurements.

The shortest distance method: defines the distance between two elements with the smallest distance in two classes as the distance between two classes, that is, the distance between two classes.

$$D_S(C_a, C_b) = \min\{d(x, y)|x \in C_a, y \in C_b\} \tag{8}$$

The longest distance method: defines the distance between the two elements with the greatest distance in two classes as the distance between two classes, that is, the distance between two classes.

$$D_L(C_a, C_b) = \max\{d(x, y)|x \in C_a, y \in C_b\}$$

Central method: if a cluster C_i, $\bar{x}_i$ represents the central point, x is a point in the class, that is, $x \in C_i$. The number of data in C_i is represented by n_i, so the center of class C_i can be defined as:

$$\bar{x}_i = \frac{1}{n_i} \sum_{x \in C_i} x \tag{9}$$

The distance between classes of two classes is defined as the distance between classes of two classes.

Then the class distance between C_a and C_b is:

$$D_C(C_a, C_b) = \mathrm{d}(r_a, r_b) \tag{10}$$

Class average method: it defines the distance between any elements in the two classes as the distance between classes.

$$D_G(C_a, C_b) = \frac{1}{mh} \sum_{x \in C_a} \sum_{y \in C_b} \mathrm{d}(x, y) \tag{11}$$

3 Implementation model of financial management information system based on Data Mining

3.1 Thinking and structure of model

At present, the financial analysis work of many large enterprises or group companies in our country still stays on manual analysis, or calculates some financial analysis indicators through financial software. Such traditional financial analysis models have the following main disadvantages:

1. The backward technology results in the analysis results being fixed reports and formats, which can not be flexibly presented according to the needs of decision-making levels at all levels;
2. Financial analysis results often stay in a certain time or space, lacking continuity, and can not dynamically reflect a problem;

3. Because of the huge workload, the response speed is slow and the timeliness is poor.
4. The scope of analysis is narrow and can not be combined with other departments of the enterprise for comprehensive analysis. Therefore, the establishment of an effective financial management information system is of great significance.

According to the characteristics of data warehouse and data mining technology, a frame structure is built here, as shown in Fig. 4.

Based on this framework, we will apply data warehouse and data mining technology to financial management, and build an excellent financial management information system. In order to avoid the drawbacks of the traditional financial system, a financial management information system model based on data mining and data warehouse technology is established. The structure of the model is shown in Fig. 5. From the structure diagram, we can clearly see the analysis basis of the model. For the data mining model based on data warehouse, the system is divided into two subsystems: financial information processing subsystem and financial management decision-making subsystem.

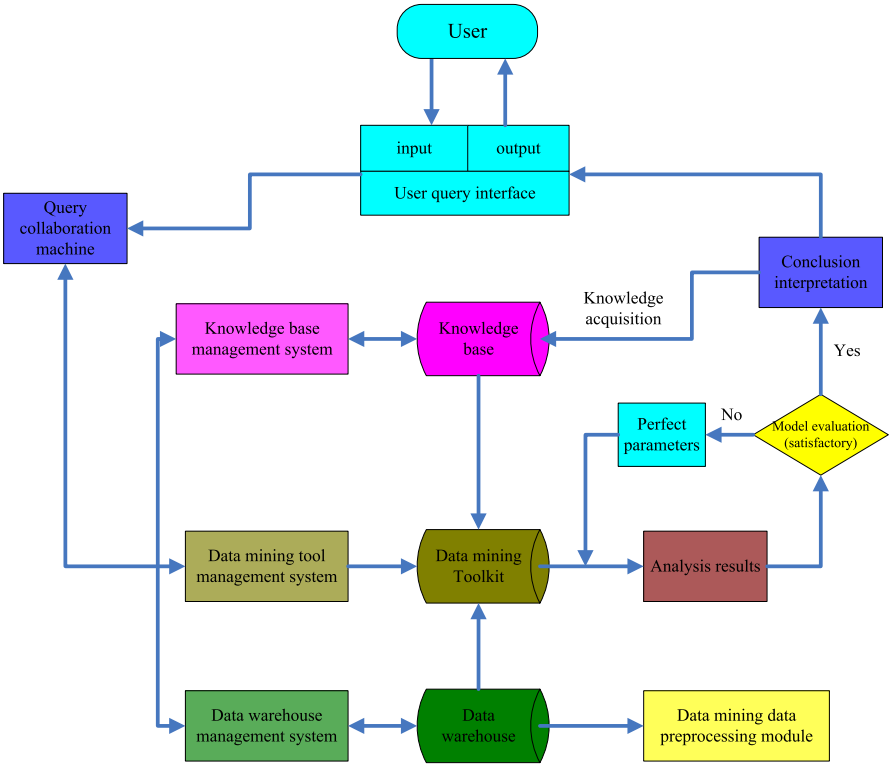

Fig. 4 A data mining framework based on data warehouse

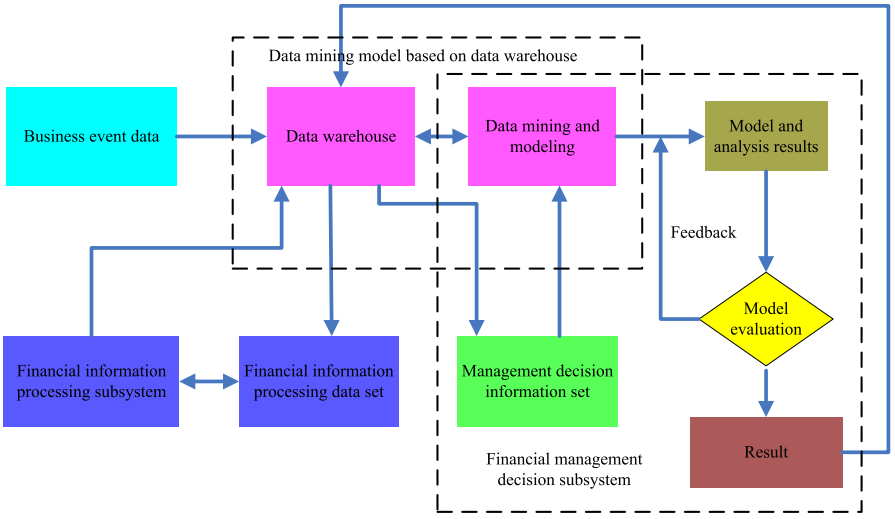

Fig. 5 Financial management information system model based on data warehouse and data mining

3.2 Related data preparation

1. Data acquisition

   Data in data warehouse comes from many kinds of economic business of enterprises. Under event-driven architecture, business events are taken as atomic units to collect data. The method is as follows: (1) Establish a model to describe the basic characteristics of business processes and events, so as to determine the entities involved, and then use the entity relationship (ER) model graph or object-oriented (OO) model to represent entities and objects, as well as their meanings and relationships with each other. Entities can be events, people, objects, locations and concepts. Each entity is described by a finite set of attributes (characteristics). (2) use entity relationship (ER) model as guidance to create entity data structure table and identify key attribute of table. (3) storing tables in the data warehouse and realizing the connection between tables through the primary key of the table. (4) When economic business occurs, the attribute values of entities are input into physical data tables by salesmen, bar code input machine or other data input means. (5) To standardize, filter and match these business data, purify and label the time stamp of data, and integrate the data from these heterogeneous information sources tightly to meet the needs of data mining.
2. Data selection and classification processing

   The data in the data warehouse are selected and classified according to the needs of financial information processing and financial management decision-making. The data needed in the two fields are separated to form the data mart of financial information processing and financial management decision-making. Based on the data mart of financial information processing, the driver program of financial information accounting is developed to generate various visual infor-

mation reports according to the needs of users. Because financial management decision-making needs to make use of the information provided by financial information processing, the data processed by the financial information processing subsystem should be returned to the data warehouse besides being provided to the information users. After data warehouse processing, it joins the financial management decision-making data mart. On this basis, using data mining technology, it establishes the management decision-making driver to provide decision support for enterprise managers.

3.3 Financial information processing subsystem

Financial information processing involves a great deal of business content. At the same time, due to the limited space, this paper takes the sales/receipt business process as an example to illustrate its implementation under the event-driven architecture. In order to describe the relationship among events, resources, locations and participants in the sales/receipt business process, and the impact of business rules and characteristics controlling the business process on the relationship between entities, a general REAL model is presented, as shown in Fig. 6.

According to Fig. 6, key attributes of various basic tables, supplementary tables and tables can be established. Event basic tables include customer order table, removal inventory table, shipment commodity table and receipt table. Incident supplementary tables include order/inventory, removal/inventory, shipment/commodity, receipt/shipment, receivables/shipment, receivables/receivables, and resource tables include inventory, cash and receivables. Participants' tables include employee list, customer list, cashier form and bank form. The data items of each table (due to limited space, the content of data items is slight) are composed of attributes describing

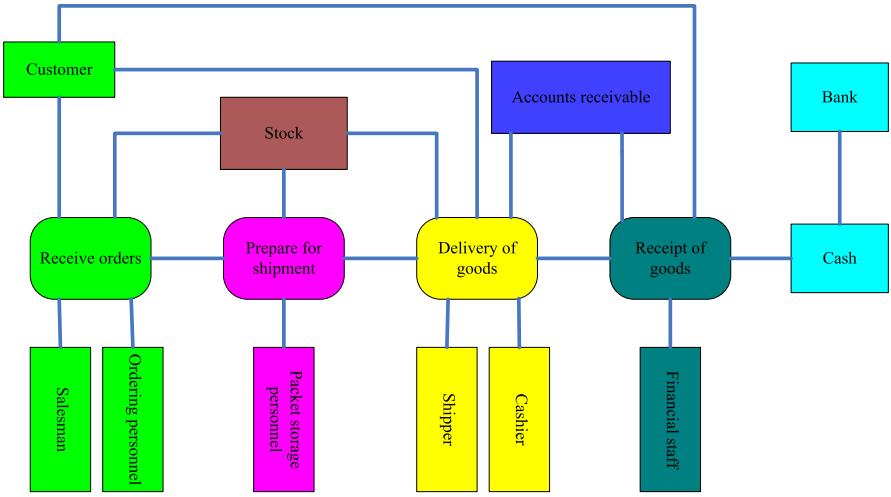

Fig. 6 REAL model of sales/receivables business

the entities of the table. By connecting these tables with the primary keys of each table, an output view without quantity restriction can be developed.

Assuming that starting with recording a cash, the cash income comes from the customer's payment of the invoice, the driver must access the cash receipt form, the customer form, the employee form, the cash form, the shipping commodity form and the receipt/shipment form in order to record the cash income. The procedure adds a record reflecting the cash receipt to the cash receipt form, and adds a record to the receipt/shipment form for each invoice corresponding to the cash receipt. The data needed to generate sales invoices are stored in the shipping commodity table, inventory table, employee table, customer table, shipping/merchandise table. The data needed to generate the vouchers can be easily accessed by connecting keywords of the data tables. The generation of sales invoices without cash receivables and the record of accounts receivable are similar to those mentioned above. Under the event-driven architecture, the balance of accounts receivable in the general ledger can be obtained by subtracting the total sales from the shipping commodity statement from the total customer payments received in the receipt statement. For the detailed accounts receivable, the required information can be obtained by formatting the output according to the customer; for the total sales in the profit and loss statement, the user only needs to add up the sales in the shipping commodity table with the specified starting and ending time. For the data of items in the balance sheet, it is only necessary to query the balance of the resource table (such as the inventory table and the accounts receivable table) on a specific date. It should be pointed out that the accounting data processing scheme under the event-driven architecture is only a change of data collection, storage and processing methods, and does not violate the accounting standards and systems that regulate how to report financial statements information.

4 System evaluation and experimental analysis

In the test of three common data sets, the generation rules of common data sets and the corresponding attribute information are introduced. Then, different common data sets are selected on the experimental platform, and the number of samples of a single tree is input. After running the program, the test results are obtained. When

Table 1 Comparison of test results for common datasets

Dataset/project	MONKS problem		ABC alphabet data set		A–E alphabet data set		SEA data set	
Training sample size	319		1592		2575		50,000	
Test sample size	135		802		1310		10,000	
Sample number of single tree	100	200	100	200	200	400	500	1000
Training accuracy rate	95.20%	94.30%	96.50%	97.40%	86.70%	87.30%	99%	98.60%
Test accuracy	66.30%	69.20%	94.50%	95.60%	80.30%	80.40%	90.60%	90.70%
Run time (SEC)	0.75	0.91	32	28	108	119	192	197

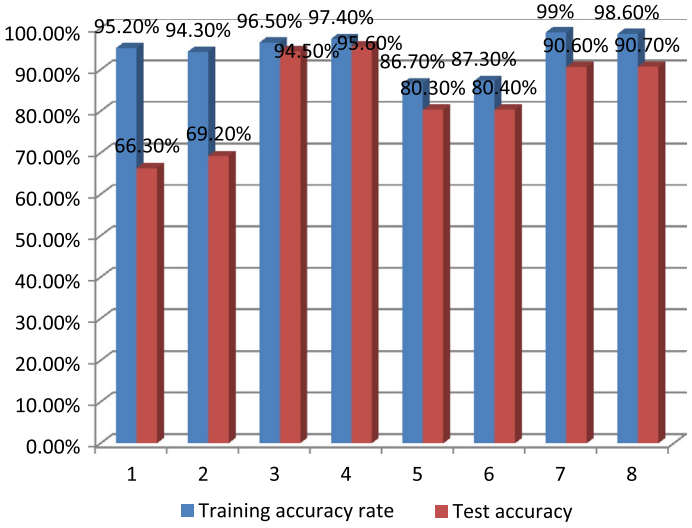

Fig. 7 Correct comparison chart

testing with MONKS-1 data set, the performance of integrated classifier is not very good, but for other data sets, the performance of integrated classifier is very good. As shown in Table 1 and Fig. 7, it can be found that the average training accuracy and test accuracy remain basically unchanged for each data set with the same number of training samples and test samples when selecting different sample numbers of a single tree. It shows that the performance of the system is relatively stable, and the running time will increase with the increase of the number of samples. The running time of the MONKS data set with the least samples is only 0.7 s, and the longest time appears in the test of SEA data set. Because SEA data sets have the largest number of samples, the structure of decision tree is more complex, but more than three minutes is acceptable here.

In the test of three financial data sets, firstly, some data extracted from the original financial system are enumerated and sorted out. Then, different financial data sets are selected on the experimental platform, and the sample number of a single

Table 2 Comparison of test results for financial data sets

Dataset/project	Goods and services expenditure data set		Payroll and welfare expenditure data set		Traffic cost data set	
Training sample size	736		754		693	
Test sample size	300		300		300	
Sample number of single tree	10	20	10	20	10	20
Training accuracy rate	96.30%	94.50%	99.1%	99.2%	96.3%	96.4%
Test accuracy	89.40%	90.20%	96.1%	92.7%	89.5%	88.7%
Run time (SEC)	4.5	6.8	1.9	1.3	3.9	6.4

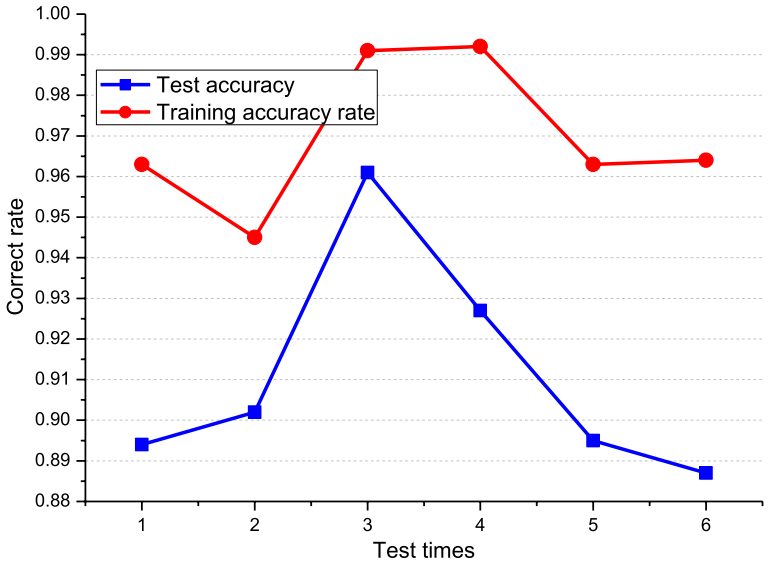

Fig. 8 Correct comparison chart

tree is input. After running the program, the test results are obtained. As shown in Table 2 and Fig. 8, it can be seen from the observation that the number of training samples is the same as the number of testing samples. Although the number of single tree samples is different, the average training accuracy and testing accuracy are relatively high, basically between 88% and 99%. It shows that the performance of classification prediction is quite good, and the running time is very short. The data set of wage and welfare expenditure has fewer fields, fewer branches of decision tree and simple structure, so it runs fast with only 1.2 s. The data sets of goods and services expenditure have relatively more fields, more branches of decision tree and more complex structure, but the longest running time is only 6.9 s, which is acceptable.

Generally speaking, both public data sets and business data sets can run smoothly with less time and no program errors. The test results (except MONKS-1 data set) basically meet the pre-requirements, especially on the test financial data set, the accuracy rate has reached more than 80% and maintained a stable state, indicating that the algorithm is suitable for the use of peacetime work, and can play a supporting role in the financial work of the state tax. This test is only part of the data set, there are a large number of financial data sets have not been tested, which subjects of data analysis can be used in practical work, need to be tested before making conclusions.

The algorithm also has the following disadvantages:

1. Because when a single sample arrives, the integrated classifier is updated, and the establishment of a single tree is time-consuming. When a large number of samples arrive, the operation efficiency decreases significantly, resulting in a decline in

the speed of operation. Moreover, it is only suitable for data sets that can reside in memory. When the training set is too large to reside in memory, the program will interrupt execution.
2. Because ID3 itself is a non-incremental learning algorithm, at the same time, when calculating information gain, it tends to choose attributes with more values, some of which are unreasonable, and other commonly used algorithms can be considered to replace them.
3. In the process of constructing the decision tree, it is necessary to scan and sort the data sets in order many times, which leads to inefficient algorithm.
4. A lot of manual judgement is needed, which is not intelligent enough. Noise is more sensitive, noise refers to the training class attribute class error or missing data.
5. The experimental platform has not been fully integrated into the financial management system. It is not very convenient to use it by calling the interface program.

5 Conclusion

This paper presents a framework of data mining and clustering analysis model for financial management information system based on improved genetic algorithm. Combining with relevant knowledge of financial analysis, an event-driven architecture is adopted to build a financial management information system model based on data mining technology. It can not only bring the ability of data warehouse and data mining technology into full play in decision support, but also make full and effective use of enterprise financial information and non-financial information. But in fact, the existing financial information system is basically based on the traditional financial information system architecture. At the same time, the application of data warehouse and data mining technology in the field of financial analysis is still in the initial stage, so the main purpose here is to build a train of thought. Three common datasets and three national tax financial expenditure datasets are selected to test the algorithm on the experimental platform. The test results show that except MONKS-1 dataset, the accuracy of the other datasets is more than 80%. Moreover, in the case of different sample numbers, the accuracy rate keeps relatively stable or rising, which shows that these data sets are well adapted to the classification algorithm and play an assistant role in the actual financial work.

References

Bider I, Jalali A (2016) Agile business process development: why, how and when—applying Nonaka's theory of knowledge transformation to business process development. Inf Syst e-Bus Manag 14(4):693–731

Breuker D, Delfmann P, Dietrich HA et al (2014) Graph theory and model collection management: conceptual framework and runtime analysis of selected graph algorithms. Inf Syst e-Bus Manag 12(1):69–106

Cavuoti S, Garofalo M, Brescia M et al (2014) Astrophysical data mining with GPU. A case study: genetic classification of globular clusters. New Astron 26(1):12–22
Chang HH, Lin CL (2015) A novel information technology of load events detection for the energy management information systems. Inf Syst e-Bus Manag 13(2):289–308
Chi Q, Fu XL, Pan YN et al (2014) The optimization model of in job-shop scheduling problem with alternative machines based on improved genetic algorithm. Appl Mech Mater 607(4):569–572
Das SP, Padhy S (2018) A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting. Int J Mach Learn Cybern 9(1):97–111
Engel R, Krathu W, Zapletal M et al (2016) Analyzing inter-organizational business processes. Inf Syst e-Bus Manag 14(3):577–612
Guo W, Xu T, Lu Z (2016) An integrated chaotic time series prediction model based on efficient extreme learning machine and differential evolution. Neural Comput Appl 27(4):883–898
Iquebal AS, Pal A, Ceglarek D et al (2014) Enhancement of Mahalanobis–Taguchi system via rough sets based feature selection. Expert Syst Appl 41(17):8003–8015
Kaiser C, Kröckel Johannes, Bodendorf F (2013) Simulating the spread of opinions in online social networks when targeting opinion leaders. Inf Syst e-Bus Manag 11(4):597–621
Ke Q, Shaofei W, Wang M, Zou Y (2018) Evaluation of developer efficiency based on improved DEA model. Wirel Pers Commun 12(4):3843–3849
Kridel D, Dolk D (2013) Automated self-service modeling: predictive analytics as a service. Inf Syst e-Bus Manag 11(1):119–140
Macas M (2014) Binary social impact theory based optimization and its applications in pattern recognition. Neurocomputing 132(7):85–96
Mikalef P, Pappas IO, Krogstie J et al (2017) Big data analytics capabilities: a systematic literature review and research agenda. Inf Syst e-Bus Manag 2:1–32
Ng KH, Khor KC (2015) StockProF: a stock profiling framework using data mining approaches. Inf Syst e-Bus Manag 15(1):1–20
Singh P (2017) A brief review of modeling approaches based on fuzzy time series. Int J Mach Learn Cybernet 8(2):397–420
Tong X, Lin J, Ji Y et al (2017) Global optimization of wireless seismic sensor network based on the Kriging model and improved particle swarm optimization algorithm. Wirel Pers Commun 95(3):1–20
Wu S (2015) A traffic motion object extraction algorithm. Int J Bifurc Chaos 25(14):Article Number 1540039
Wu S, Wang M, Zou Y (2018) Research on internet information mining based on agent algorithm. Future Gener Comput Syst 86:598–602
Xiao Y, Liu JJ, Hu Y et al (2014) A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting. J Air Transp Manag 39(39):1–11
Xie Y, Takala J, Liu Y et al (2015) A combinatorial optimization model for enterprise patent transfer. Inf Technol Manag 16(4):327–337
Younus ZS, Mohamad D, Saba T et al (2015) Content-based image retrieval using PSO and k-means clustering algorithm. Arab J Geosci 8(8):6211–6224
Zacharewicz G, Diallo S, Ducq Y et al (2016) Model-based approaches for interoperability of next generation enterprise information systems: state of the art and future challenges. Inf Syst e-Bus Manag 15(4):1–28
Zoet M, Versendaal J (2014) Defining collaborative business rules management solutions: framework and method. Inf Syst e-Bus Manag 12(4):543–565