Study on Corporate Financial Data Analysis System Based on Data Mining

Pengwu Wang

School of Accounting, Harbin University of Commerce, Harbin, China, 150028

346591653@163.com

KEYWORDS: Data Mining; Financial Data; Analysis System

ABSTRACT: With the development of economy and technology, data mining which is based on databases and business intelligence has been developed and widely applied to all fields, financial field included. It can explore hidden, useful information to help decision makers to search for the relationship among data and find out what has been ignored. Compared with traditional financial analysis, data mining can deal with massive financial data, to help the company's investors and policy makers to have in-depth understanding of the company's financial situation, and make right decisions. Here we design a financial data analysis system based on data mining model, and we conduct a comparative test by Logistic regression algorithm and decision tree algorithm, and the results show that using data mining algorithms to predict ROE business is feasible.

I. Introduction

To use data mining method for company financial data analysis, the key is data processing. Daily date produced by modern enterprise is quite massive, especially financial data. This requires us to build a model to simulate the process of financial data analysis. Data mining methods used in the financial analysis of enterprises is very appropriate.

The traditional financial analysis methods of enterprises usually focus on multivariate analysis, that is, make comprehensive analysis on the financial situation of enterprises from multiple perspectives, in order to get a more comprehensive and more reasonable analysis results. However, with the continuous development of computer technology, enterprises have accumulated and collect more and more data, and enterprises are facing more and more fierce competition, and the limitations of traditional financial analysis has become increasingly prominent, so data mining applied to financial analysis becomes more and more urgent and necessary. Nowadays, with the development of the domestic stock market in recent years, more and more experts and scholars have begun to study how to use data mining techniques to analyze financial data of listed companies.

II. Overall Design

A. System Architecture Design

The financial data analysis system we designed is based on data warehouse and data mining technology to achieve massive business data storage; it uses parallel processing technology to deal with complex query request services, to realize the decision support query optimization, and meanwhile support multidimensional analysis of query patterns.

The system is based on the architecture of data warehouse- multidimensional data processing-data mining. The overall architecture of the system is divided into three layers: data acquisition layer, data storage layer, data analysis and presentation layer, as shown in Figure 1:

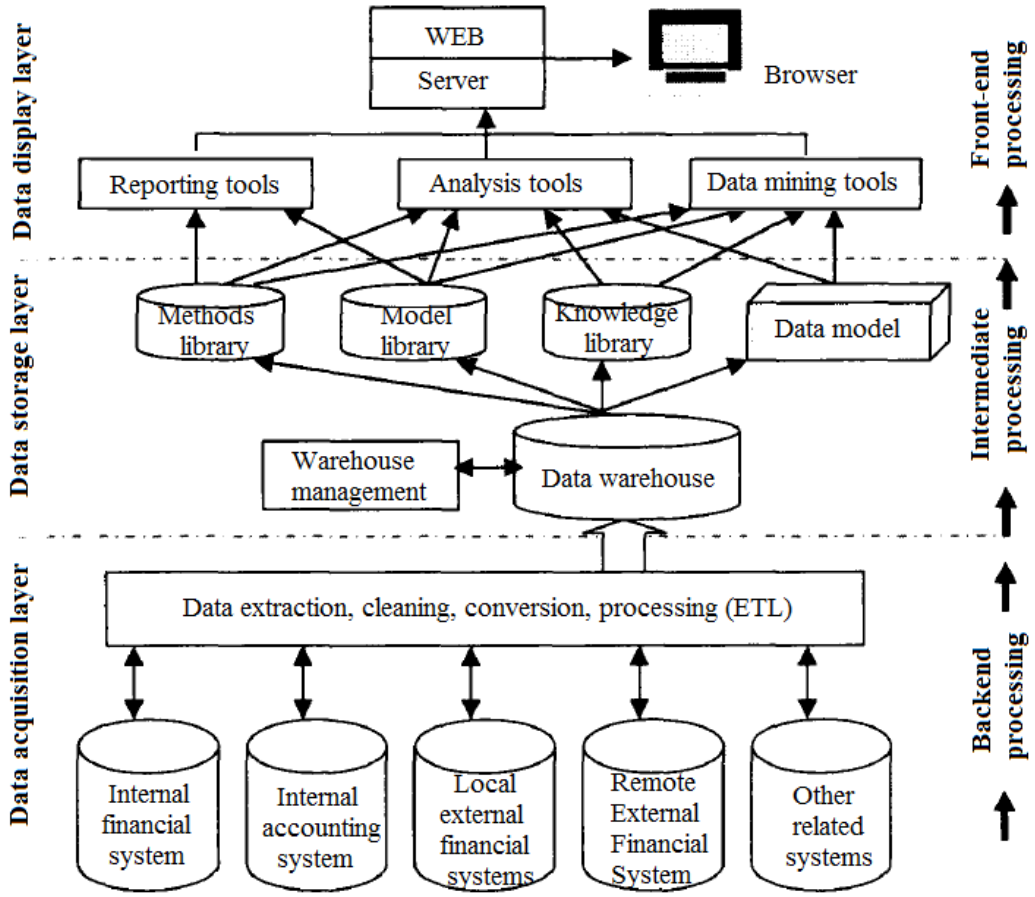

Figure 1. Overall system structure

The system realization structure includes three main bodies. The first one is combination of model library system and database system, which is the basis of financial data analysis, providing quantitative analysis information for decision making problem; the second is data warehousing and OLAP, which extracts comprehensive data and information from the database; the third is combination of expert systems and data mining.

B. System Functional Requirements

(1) Financial data capture

Financial data of listed companies are from the financial statements released by them. However, all this information is published in the form of non-edited documents, so we need to grab the financial data from the financial statements in these documents, and store in database, which could be edited, queried and analyzed.

a) Classify Listing Corporation's financial statements in terms of Stock exchange, time, report type, etc.

b) Financial statements are mostly PDF format. Transform the PDF format into html format, so that the structure elements could be identified and extract available financial information.

c) Use htmlParser to resolve the financial reports which have been transformed into HTML format, and then find its corresponding financial statements and extract financial data from it.

(2) Financial data analysis

The financial data extracted from financial statements by the system is noisy data, and it is not stored according to business theme, so it cannot be directly used for data analysis. First of all, we need to deal with the data by empty value processing, standardization of data formats, correctness verification and other operation, and then establish data warehouse in terms of subject classification according to the processed data. On basis of the data warehouse, carry out query and analysis of data based on business. This function mainly consists of two parts:

a) Establish data warehouse. After data cleansing, establish data warehouse in accordance with the theme of analysis.

b) Based on the data warehouse, carry out data analysis in accordance with business requirements, so as to dig out the valuable information hidden in the data.

(3) Visual display

Financial data analysis is the core function of the system. Traditional financial data analysis is manual analysis based on statistical methods, which is inefficient and prone to error. In this study, we would calculate the basic indicators of the financial situation of the enterprise according to the data from the crawled financial statements, and then use the data mining algorithm to classify the listing corporation, and analyze the factors that affect the development of enterprises from the perspective of Industry, business scale, time, etc.

Return on net assets is a core index reflecting the profitability and management level of listing corporation. The higher this index value, the higher the return on investment will be. Therefore, this system takes the return on net assets as the index to evaluate the investment value of the enterprise. When the rate of return on net assets of the enterprise is lower than the deposit interest rate of the current year, this shows that it is not worth the investment the enterprise. The user can set that when the net assets higher than the standard value, the system prompts the user to pay attention to this company, and meanwhile it will list the company's financial indicators and that of the same industry average.

Finally, system shows the analysis result information and the basic information of the listing corporation through the visual way. The system can support a variety of ways to show, including line chart, bar chart, pie chart and other chart, and allow user to check the information about the listing corporation from the industry, financial projects, financial indicators and other statistical items.

III. System Implementation

A. Data Mining Model

From the perspective of profitability, earnings quality, solvency, operational capacity and development, based on Logistic regression model and decision tree model, we conduct comparative test. The establishing processes of these two algorithms models are as follows:

For Logistic regression model, we use backward stepwise method to select variables to entry the model. That is, under the premise that all the independent variables are contained in the model, remove the independent variables which do not conform to the requirements. The specific steps of screening are as follow:

(1) All variables enter in the model have been defined;

(2) For all variables, calculate the Wald test value, and obtain the corresponding p value;

(3) Find out the property of the maximum value of p. If it is larger than the defined significant level value, then remove this variable; if there is no variable to remove, then the screening process terminates;

(4) Go back to step (2) and continue to remove.

For decision tree algorithm, we should analyze the future investment value of enterprises, especially the future rate of return on net assets. Establishment and application process of decision tree is as follow:

(1) According to the classification standard of each index, compile and categorize the ROE, net profit growth rate(NPGR) and sales growth rate(SGR) in previous years, and save them into the database;

(2) According to historical financial data, perform decision tree algorithm. According to the classification criteria of ROE, generates the appropriate decision tree, and store it in the database with a certain form;

(3) After the completion of the above steps, with listed company's financial data in current few years, it can be showed according to the generated decision tree, so that we can carry out evaluation on listed company's EOR in the coming year, to assist investment decisions.

B. Experimental Results

We adopt Logistic regression algorithm and decision tree algorithm for comparison test. In our model, the modeling data are crawled from published financial reports in stock exchange website, that of 2009 to 2013 as independent variables, while that of 2014 as dependent variable, to predict the listed companies whose ROE is not less than 15%. Then take the financial data from 2010 to 2014 as independent variables, and ROE of 2015 as dependent variable, as test set to test the previous generation model train set.

For Logistic regression algorithm, the final result is as shown in Table 1:

Table 1. Logistic regression experimental result

Parameter	Parameter estimates	Standard deviation	Wald statistic	Degree of freedom	Significant level
Historical ROE(F1)	-13.893	3.873	12.819	1	0.000
Historical NPGR(F2)	-5.293	1.465	10.287	1	0.001
Historical SGR(F3)	-3.291	0.901	6.893	1	0.003
Intercept	2.593	1.032	5.234	1	0.019

Thus, we can obtain the prediction model by Logistic regression algorithm on listed company's financial data training set as follow:

$$\log\left(\frac{p}{1-p}\right) = 2.593 - 13.893F1 - 5.293F2 - 3.291F3 \tag{1}$$

For decision tree algorithm, by the construction of list company’s financial data training set, we have obtained the following rules:

(1) When ROE of the year is higher than 25%, then the probability will be greater than 90% that ROE of next year is higher than 15%.

(2) When ROE of the year is lower than 25%, while NPGR is higher than 45%, then the probability will be greater than 80% that ROE of next year is higher than 15%.

(3) When ROE of the year is lower than 25%, while NPGR is lower than 45% and SGR is higher 60%, then the probability will be greater than 75% that ROE of next year is higher than 15%.

For test data set, the prediction results of Logistic regression algorithm and ID3 decision tree algorithm are shown in Table 2:

Table 2. Prediction result by the two algorithms

Parameter	ROE	Prediction result		Accuracy
		More than 15%	Less than 15%	
Logistic	More than 15%	262	54	80.81%
	Less than 15%	253	1031	
Decision tree	More than 15%	297	19	86.38%
	Less than 15%	199	1085	

From the prediction results, we can see that using data mining algorithms to predict ROE business is feasible, and the accuracy is about 80%, which is a reliable data. We can also find that, decision tree is better than Logistic regression algorithm, although both of them are given a high probability value, so compared with Traditional qualitative analysis method, it has made great improvement.

REFERENCE:

[1] Codreanu D E, Popa I, Parpandel D. Accounting and Financial Data Analysis Data Mining Tools[J]. Eirp Proceedings, 2011, 6(1).

[2] Uzar C. The Usage of Data Mining Technology in Financial Information System: An Application on Borsa Istanbul[J]. International Journal of Finance & Banking Studies, 2014, 3(1):51-61.

[3] Zhu J, Shi-Jun L I. Based on Data Mining in Financial Data Analysis[J]. Computer Knowledge & Technology, 2010.

[4] Zhang B W. Data Mining Technique in the Application of Financial Information Systems[J]. Computer Knowledge & Technology, 2011.

[5] Shu J W S. OLAP-Based Multi-Dimension Analysis System for Financial Data in Construction Enterprises[J]. Chinese Journal of Management, 2005.