Intelligent Data Mining for Economic Prediction and Analysis

Zhao Yongyi
Software College
Shenyang Normal University
Shenyang, China
Michael_NEU@126.com

Pan Qiang
Software College
Shenyang Normal University
Shenyang, China
panqiangg@126.com

***Abstract*—This paper describes importance that the application of economic data in the data mining algorithm and its application, which combines with the current economic data of national macro-economic indicators, presents the data warehouse model structure and its implementation characteristics, and uses SQL Server 2005 data warehouse and data mining solutions on economic data for the application of data mining solution, system architecture, algorithms implementation, and finally discusses the application of data mining algorithms development trends and key technologies in the economic field.**

Keywords-data mining; economic prediction; dataware house; OLAP; SQL Server

I. INTRODUCTION

With the maturity of the data mining technology and widely application, especially the development in the economic field, which provided us a strong analysis and decision support tools. Currently, the economy of China develops rapidly, the economic data is growing, more and more the economic types are appearing, in order to make better use of economic data which are very useful resources, and these data must be taken with a certain degree of technical means and scientific methods to establish a reasonable and complete, standardized and scientific architecture. Therefore, how to center on economic data, analysis of data mining and decision-making has become the focus of our work, and the complete of its construction and development will be a direct impact on social development.

Microsoft SQL Server 2005 mainly included: integration services, analysis services, reporting services component. It provides users build models and innovative analytical applications, which required for various features, tools and others.

Through AMO and ADO.NET components that offered by analysis services. The system has been developed through the dynamic visual interface to create curb, dynamic add and modify dimensions (from the point of view that user analyses the fact data), satisfies the needs of different user's analysis requirement, accesses to flexible data query, gets the data query quickly, analyses business data from multi-angle. In addition, the combination of data mining algorithms will be a classic multi-dimensional data and centralized data into information and knowledge to provide users with decision support [1][2].

II. ALGORITHMS OF ECONOMIC PREDICTION

A. Linear regress algorithm

After analyses and the research on economic data, the system that towards to algorithms of linear regression not only need to support the simple linear regression ($y_t = ax_t + b$), but also to increase the regression analysis ($y_t = ax_{t-1} + b$) with a time lag of variables, as well as multi-variable linear regression ($y = ax_1 + bx_2 + cx_3 + d$), meanwhile the issue of variables lag problems must be considered ($y_t = ax_{1(t-1)} + bx_{2(t-2)} + ... + nx_{n(t-n)} + d$). System model derived from the form of expression as follow:

dependent variable = constant + coefficient* variables

The test results that derived from the system model must be unified with statistical examine of concept in the measure econometric model. It has been reserved six valid values after the decimal point. The situation of predictive value and the increase of the actual value have been described in the scatter diagram which provided by SQL Server 2005 reporting services. It facilitates the practical value and predictive value to contrast. It carries out inspection and examines the prediction of model from the intuitive model forecasts on the situation intuitively.

B. Time series algorithm

The time series components include: long-term trend (*T*) is a time series with time which gradually increase or decrease in long-term changes in trends; seasonal changes (*S*) is the time series in one year or a fixed period of time, showing the fixed rules changes; change of cycle(*C*) along with the trend line cycle changes, which likes a pendulum, also known as business cycle movement; irregular changes (*I*) is defined as the time sequence, which is the result of random factors arising from the changes. Its mixed model include: additive model, it is assumed that the time series is based on four components which derived from the sum. Long-term trend does not affect the seasonal changes. If Y is the time series, the additive model is $Y = T + S + C + I$. Multiplicative model, it is assumed that the time series is based on four components derived from multiplying and that the seasonal changes and cycle changes are the function for the long-term trends, so the model of the equation is $Y = T \times S \times C \times I$. It should be noted that due to time series prediction method did not consider the impact of external factors to highlight the time series, so there is a prediction bias of the defects. When facing with large changes in the outside world, usually, there is a greater deviation, time series forecasting method in long-term more effective than short-term forecasts predict.

C. Decision tree algorithm

The Microsoft decision trees algorithm is a classification and regression algorithm provided by Microsoft SQL Server analysis services for use in predictive modeling of both discrete and continuous attributes.

For discrete attributes, the algorithm makes predictions based on the relationships between input columns in a dataset. It uses the values, known as states, of those columns to predict the states of a column that you designate as predictable. Specifically, the algorithm identifies the input columns that are correlated with the predictable column. For example, in a scenario to predict which customers are likely to purchase a bicycle, if nine out of ten younger customers buy a bicycle, but only two out of ten older customers do so, the algorithm infers that age is a good predictor of bicycle purchase. The decision tree makes predictions based on this tendency toward a particular outcome.

For continuous attributes, the algorithm uses linear regression to determine where a decision tree splits.

If more than one column is set to predictable, or if the input data contains a nested table that is set to predictable, the algorithm builds a separate decision tree for each predictable column.

III. ALGORITHM APPLICATIONS

A. System structure

System uses Microsoft SQL Server 2005 to build data mining algorithms of the upper application, and its logic structure is divided into three layers:

- Presentation layer. The various departments, system administrators and enterprises users, according to their different permissions to complete data mining, data statistical analysis, data conversion and other related functions.
- Logic layer. It is mainly completion of users to access authentication, service mapping, integration services API, analytical services API, Data Mining API and so on, and part of its function is to serve mapping of which the main function of different roles for different users is the access to analysis services and the reporting services to provide real-time mapping. Furthermore, its purpose is that the data indicators of system were changed or strategies of statistical analysis were changed, the system without any programming can be adapted to these changes.
- Data layers. It based on the Microsoft SQL Server 2005 data warehouse to create applications for the upper relational database, data warehouse persistent data support, as shown in Figure 1.

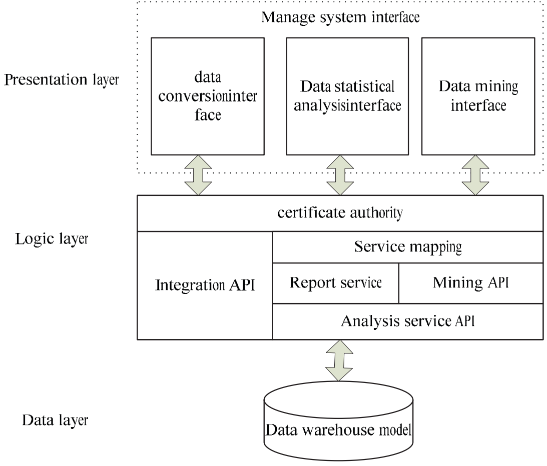

Figure 1. System logic structure

Through visualization interface, the system realizes to create cube, customize personalized multi-dimensional data dynamically, and then shows the analysis results to the end user by various forms, like the table or the figure. The basic technical scheme of the analysis and realization will be present as follows. Through the SQL Server 2005's integration services component, the system stores various forms data which were collected from the bottom to form a unified model into data warehouse that has been split in accordance with the theme of the data. In the data warehouse, the system was derived by multi-dimensional data cube model organizations from all kinds of subject data, and then conducts OLAP operations on the cube. Finally, the system forms a complex cube query and report analysis statements. Of which: data warehouse's data comes from a number of scattered OLTP database, spreadsheet data, text documents, internet data, and so on. First of all, the system through the SQL Server 2005 integration services extract data from data source to the data warehouse. During this process the various data come from sources data need to go through filtration being formed after the conversion and integration then formed data set which has consistent model in the global, and then the establishment of data warehouse on this basis[3][4]. Data warehouse organizes data in accordance with the theme to the completion of a department or organization in decision-making task of intelligent analysis. Data warehouse was built in multi-dimensional model, mainly includes star-shaped pattern and snow-shaped pattern, and contains the details of data and aggregated data. After the establishment of data warehouse, the system can not only be in the data cube based on the completion under the drilling analysis which is from a general to the specific, but also to be completed drilling analysis which is from a specific to the general.

Extract, transform and load (ETL), the data that come from the data source will be filtering, transformation and integration, and then loaded into the data warehouse. We use SSIS interface and the automatic analysis technology and the expert knowledge strategy to filter data in order to remove the dirty data in original data (such as duplicate tuple, false data, etc.), and use integration technology of database implement multi-data sources semantic integration, at the same time converse and processing relevant data that based on the system's demand, such as changes in data types, generate new fields, and so on [5].

MDX is a sentence-based and functional completeness script language, which has been used to define, use and search the data in the multi-dimensional object from Microsoft SQL Server 2005 Analysis Services (SSAS). MDX is the most important component element in the unique dimensional model method (UDM) that combines the XML with analysis (XMLA) protocol. It not only supports to search the multi-dimensional data from the dimensional model, but also has the capacity to search the basic data from relational data source or table data source.

XML for analysis SDK (XML/A SDK) includes two parts: XML for analysis provider and sample client application. The XML for analysis provider provides the capacity to access analysis data source (OLAP and data mining) on the web. The XML for analysis provider provides the unique access method to access the analysis data source trough executing XML for Analysis Specification that do not need develop client component to realize the COM's interface.

For the access to the multi-dimensional data, the system must install the XML/A SDK for the communication between the client and server.

B. Data warehouse structure of economic prediction

In this paper, the data warehouse model has been constructed out of using the Microsoft SQL Server 2005 to build the upper application of data mining algorithms, using the macroeconomic data as an example and using star-shaped data warehouse model to create the fact table and dimension table. It has been described the logical structure of deposits in the data warehouse in Figure 2. A specific value of macroeconomic indicators is stored in fact table. The area dimension table, the unit dimension table, the time dimension table and the measure value dimension table respectively store information of macro-economic indicators. The fact table joins with each dimension table by serial number, area ID, and measure value ID, and unit ID, and time ID. It ensures that analyze of fact data demands from varied lights, area dimension, unit dimension. The time dimension is a conventional dimension and the measure value dimension degree is achieved by the father-son dimension.

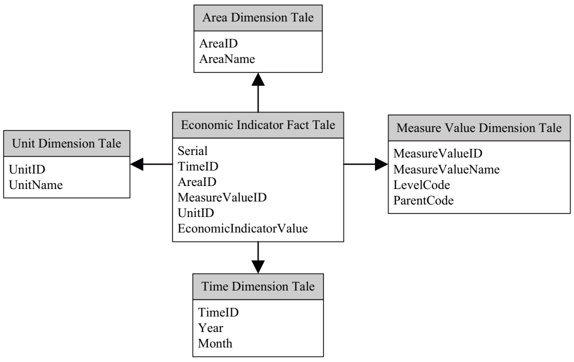

Figure 2. Data warehouse model

IV. ALGORITHM IMPLEMENTATION

In this paper, which cites the linear regression algorithm, the time series algorithm and decision tree algorithm as examples, it illustrates how to implement the system of data mining algorithms. First of all, user login in the system and enter the forecast parameter selection page. This page will provide all parameters which are used to predict: time, area, one or more national economic indicators, such as GDP, and so on. And then, user selects the model to predict, as linear regression algorithm, multi-variable linear regression model algorithm or time series algorithm. It has been described about how to data mining algorithm processing procedures in SQL Server 2005 as shown in Figure 3.

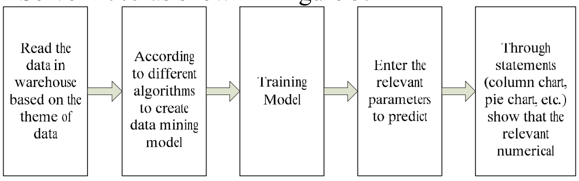
Read the data in warehouse based on the theme of data
According to different algorithms to create data mining model
Training Model
Enter the relevant parameters to predict
Through statements (column chart, pie chart, etc.) show that the relevant numerical

Figure 3. Data mining algorithm's processing flowchart

After choosing these parameters, like the time, the area, the economic indicators etc, the system accesses to the relevant data sets through the OLAP API in the data warehouse[6], then using the data sets to train the model which has been constructed already, in essence, the system uses these data to define the algorithm formula parameters. More training time would be needed if lots of indicator data have been chosen, in other words, if data size is large, it will take more training time, if data size is little, it will take less training time. These data sets were stored in a data table which is designated and associated with the model to ensure that direct invoke after the training model to reduce the direct data warehouse invoke time complexity, at the same time, in order to control analysis, it kept the predictive value obtained from a predict parameters. Once the model trained can be direct invoked and then predicted by API which was provided by SQL Server 2005 and prediction. The system must maintain the relationship between model name and data sheet, which stores the training data before each model was created dynamically, and then create table space which stores training data, for increasing the contrast with the model name of the relationship.

These relationships were stored in table, these relationships are stored in a relational table, in order to facilitate modify, and delete, such as management in the future after training of the model. For the forecast data, the system can specify the model name and put data into a specified algorithms formula. Finally, the system can get the prediction result and statements by specifying report to the user [7] [8].

V. ALGORITHM PERFORMANCE OPTIMIZATION

We found that there are two problems around the training time. The first is that economic data is large; the second is that data relationship is more. For the second problem, SQL's execute time is less than MDX's. So we create the temporary table to store the temporary data. At the same time we use the simple temporary table structure to reduce the data relatedness. In the same data size, the process time is less than before.

VI. CONCLUSION

This paper analyses the general process of data mining, and introduces a number of data mining algorithms which are commonly used in economic analysis. Those are linear regression algorithm, multi-variable linear regression algorithms, decision tree algorithm and time series algorithm. And then the system has been implemented in Windows Server 2003 + Visual Studio 2005 + SQL Server 2005 environment to achieve a linear regression (one variable, multi-variable) algorithm, decision tree algorithm and time series algorithm.

After the research on the linear regression analysis algorithm, multi-variable linear regression analysis algorithms, decision tree algorithm and time series algorithm, the system constructs the actual national macro economic data warehouse model and the implement data predictive function by linear regression algorithms, decision tree algorithm and time series algorithm, and the usage of the country's economic data warehouse model and micro economical data.

The linear regression algorithm, decision tree algorithm and the time series algorithm are commonly used in economic analysis. Using SQL Server 2005 to realize the predictive analysis is an experiment that explores the combination scheme about data warehouse and substantive application, like SPPS or EView's economic statistics function, etc, large-scale analysis software, of which the data warehouse support is limited, but has wealthy resources of algorithms to evaluate, and the SQL Server 2005 is inadequate in this area. Next research will focus on how to study of error analysis algorithm to ensure the feasibility and credibility based on SQL Server 2005 data mining application programming interface.

REFERENCES

[1] Zhaohui Tang, Jamie MacLennan. Data Mining with SQL Server 2005. Beijing: Tsinghua University Press, 2007

[2] Wang zheng, Li jiaxing. SQL Server 2005 practical guide. Beijing: Tsinghua University Press ,2006:146-150.

[3] Zhang bo, Chen dingfang, Zu qiaohong. Data Mining System Design Based On SQL SERVER 2005. Journal of Hubei polytechnical University, 2007,(03)

[4] TANG Z H, MACLENNAN J. Data Mining with SQL Server 2005. Indiana:Wiley Publishing,Inc,2005:344-366

[5] Yike Guo, Robert Grossman. High Performance Data Mining: Scaling Algorithms, Applications and Systems. Germany:Springer , 2001:1-56.

[6] Hanxuemei. Time-serious mining and prediction research. Zhejiang: Zhejiang University,2006.

[7] Renrong,Wanglunjin. SQL Server 2005 Data mining API Technology Analysis And Real Application. Ningxia Engineering,2007: TP391:19-21

[8] Xiejiabin. The research of method of prediction decision-making based on data mining. Jinan Univeristy, 2007.