Development on Intelligent Data Mining System in Oil Logging

Kewen XIA, Zhe ZHANG*, Yatong ZHOU, Suicheng SI

School of Information Engineering, Hebei University of Technology, Tianjin 300401, China

Abstract

In order to solve the problems, such as the very complex development process and relatively few researches on visual data mining technology in the system software of oil logging interpretation. A set of intelligent data mining system in oil logging is developed based on the Visual C++ and intelligent data mining technology, the main design on system includes data layer, logic layer and presentation layer, and the logic layer includes data preprocessing and intelligent mining technology. In addition, On the VC++ platform, the system can realize the visual functions, such as data management and graph and image data processing. The system can also be embedded with the Matlab toolbox to complete intelligent data mining. The actual application shows that the intelligent data mining system possesses friendly interface, high visualization degree, strongly expansibility and good mining effect.

Keywords: Software System; Intelligent Data Mining; Oil Logging

1 Introduction

In the field of data mining, there are a lot of data mining system widely in application at present, such as Intelligent Miner, DBMiner and so on, based on lots of framework and Algorithmic Modeling Technology [1], however, in the field of oil logging, many companies still stay the application stage with traditional data mining technology, and there are some problems in current oil logging software system, such as the very complex development process, relatively few researches on visual data mining technology, and so on.

Therefore, this topic especially discusses an intelligent data mining system for oil logging based on intelligent data mining technology and VC++ technique. In accordance with the visual data mining technology in the show layer, the development of software system can be realized by using Visual C++ programming method combined with database SQL SERVER 2005 and Matlab7.0, which can meet the need of the actual logging data mining as well.
2 System Framework Design

Due to the high requirement of interface engine and diversity data source in data layer [2] for the data mining tools, VC++ technology could be selected as the basis of the system development platform, for its powerful function library and wide range of interface support.

In the aspect of framework selection, the superiority of C/S framework is available in file management, sharing server pressure and dealing with complex business process [3], and the B/S framework possesses superiority in framework hierarchy, dynamic flexibility and data sharing. For the oil logging data is increasing in massive trend, and the calculation of the mining algorithm spends costly, the selection on the centralized C/S framework is more appropriate to this system.

In the aspect of data layer interface, it uses the API interface closely connected the unity of the C/S framework, i.e., ADO (ActiveX Data Object) technology connects database SQL SERVER 2005 to solve the calculation cost by mixing Matlab programming in background.

In the aspect of oil logging mining structure, there are three layers, such as data layer, logic layer and performance layer in the mining structure [4], as shown in the Fig. 1. The logic layer includes data pretreatment, data modeling and intelligent optimization, etc. For there are some definite requirements in the system, the waterfall model of software is adopted in software design. In the aspect of system function, the system includes user module, basic function (presentation layer) and intelligent algorithm.

3 Description of Intelligent Mining Technology

Data mining is the core of knowledge discovery link, i.e., through the methods of data pretreatment and modeling analysis to complete the purpose of extracting useful information or mining task [5]. what’s more, in order to achieve better calculation result, intelligent optimization is to use intelligent algorithm to optimize model structure or parameter.
3.1 Data preprocessing

Before mining the data, it is necessary to complete the logging dirty data processing so as to prepare for modeling by collecting all kinds of data description (such as mean value, statistic such as standard deviation) and data distribution. The data preprocessing mainly includes the data cleaning, data conversion, data normalization and attribute reduction, etc.

(1) Data cleaning. Data cleaning is to solve the problem of missing value in data, redundancy, data value, definition does not agree, outdated data, etc. Logging signal includes sound, electricity, nuclear, and many other signals, the various logging signal collected usually includes some noise, such as random noise, singular value or salt and pepper noise, etc. In accordance with the experience, the signal of acoustic slowness all apply to 5 point filtering, and the way of natural gamma ray and spontaneous potential signal must be decided by some factors such as the different characteristics of the oil well and noise intensity, so data cleaning module can be divided into three parts, there are eliminate singular value, missing value up, random noise processing (using 5 points, 7, 9, 11 point average filtering method).

(2) Data conversion. Due to the relatively simple mathematical methods involved in the data transformation, from the angle of reducing ”fat client” pressure and making program execution efficient, the function of data transformation is realized by the stored procedure, which integrates SQL statements and statement of process control.

a. Data sampling. It can be realized by using division method on ID number of table from database;

b. Data Changing. When the data changes is slow and not obvious, can use the exponential transformation formula, on the contrary, when the data grows fast, the logarithmic transformation does work.

c. Normalized. For the convenience of sample learning and prediction data for logging data, the steel quantity of all kinds of logging data has been normalized, which make the number in the range of 0 to 1., as shown in Eq. (1). It needs the way by calculating the extremum and rounding operations.

\[ f(X) = \frac{X - X_{min}}{X_{max} - X_{min}} \]  

(1)

In the light of the data with tremendous changing trends, such as resistance rate and other physical quantities, we could carry on the logarithmic transformation firstly and then make the normalization.

(3) Attribute reduction. Attribute reduction can eliminate redundancy attribute existing in the knowledge, and solve the minimum attribute reduction, which belongs to NP problem, people often use the heuristic information (such as the attribute significance) to simplify calculation in order to find the optimal or suboptimal reduction, in addition, another attribute reduction based on evolutionary computation (such as genetic algorithm, particle swarm optimization (PSO) algorithm), can carry out the mass data processing, the construction of fitness function plays an especially important role in it particularly.

If there are massive amounts of data, we can deal with sample extraction firstly, and then reduction the samples attribute.
3.2 The mining technology based on the classification and prediction

There are a lot of methods in the establishment of the data mining model, such as classification and prediction, clustering analysis and association rules and evolution analysis, characterization and distinction, and so forth [5].

Among the classification and prediction of mining technology, support vector machine (SVM) method outshine others, it is based on structural risk minimum principle. We take advantage of this thought to construct classification hyper-plane with maximum margin and decision rule, which inputs vector through mapping into a high dimensional space by specific kernel function. And SVM can be widely applied to the classification and regression estimation [6, 7]. Least squares support vector machine (LS-SVM) is also widely used in oil and water bed identification. Among them, the relevance vector machine (RVM) [8] also belongs to the SVM, its training is in the Bayesian framework based on active relevant decisions theory in the structure of the priori parameters for removing unrelated point, so it gets the purpose of sparse model, and it reflects the core characteristics of the data. Therefore, this article mainly uses the SVM method to establish the data mining model.

In the realization of the algorithm, the toolbox such as C++, Matlab has been popular, for instance, libsvm, svmlight and torchsvm are the most typical vehicle for the SVM toolbox. SB1 toolbox can solve the RVM problem in classification and prediction. However, the interface connected with data source about toolbox data is the key technology of embedded system. This mining system can use various versions of mature software tools such as mixed programming combined with Matlab toolbox and VC++ [9], platform interaction could be completed with Matlab. Implementation steps are as follows.

Step 1 Read data from database on VC++ platform;

Step 2 Put data into data storage format of Matlab, such as TXT, MAT format file;

Step 3 Trigger M file operating Matlab in background on VC++ platform with engine method;

Step 4 Convey parameters of logging results back to the VC++ system from Matlab.

3.3 Intelligent optimization

Data mining algorithm itself cannot be developed completely by once, we need constantly updating and optimizing the algorithm to meet the actual requirement [2]. Therefore, this system mainly uses the modern intelligent optimization method, such as PSO algorithm [11] and improved algorithm, quantum PSO algorithm (QPSO) [12], quantum culture PSO (QCPSO) algorithm, and semi-definite programming (SDP) [13] algorithm and second-order cone programming (SOCP) algorithm, etc. Among them, the PSO and its improved algorithm has mainly used for parameter optimization of mining model; SDP and SOCP have mainly used for judging the effectiveness of kernel function parameters, and use better nuclear matrix with combination of effective parameters to improve SVM model classification accuracy. The SDPT, ScDuMi, sdpam, and so forth, can be called in Matlab.
4 Realization of the Presentation Layer

The visualization technology of presentation layer is that using computer graphics and image processing technology converts data into the format of graphics or image, which can displays and interacts with screen [14]. The presentation layer of the system software can accommodate logging interpretation, meanwhile, complete the function such as query of basic information of well logging, logging image and loaded format conversion. Design can be divided into two big modules as data management, graph and image data processing.

4.1 Data management module

The visualization technology of presentation layer is that using computer graphics and image processing technology converts data into the format of graphics or image, which can displays and interacts with screen [14]. The presentation layer of the system software can accommodate explain information of log mining, meanwhile, complete the function such as query of basic information of well logging, logging image and loaded format conversion. Design can be divided into two big modules as data management, graph and image data processing.

1) ADO interface. Package a ADO interface classes, object of smart pointer RecordsetPtr calls dataset, and those can realize attribute value (listing) choice of logging data sheet (table name), what’s more, which can accomplish the function of reading and writing the corresponding table and listing in sections.

2) Realize data call. Implementation steps of data call are as follows. Map the corresponding control permissions and cascade relationship of logging database through the control choice; and then do logging database interaction according to the chosen of trigger events; display in performance layer at last.

Fig. 2(a) is the query results of a certain well Su in logging data management interface, whose attribute are Depth, DEN and WQ between the 3180 and 3200m interval. Fig. 2(b) is the query result diagram of a certain well A, whose limiting condition is that DT attribute value is equal to 274.4620 in the curve DT between the 3180 and 3200m interval.
4.2 Graph and image data processing module

Image data processing module realizes the logging data query by image form, including dynamic graph of well logging, function of image saving, read and format conversion, it enhances the effort of visualization, shows the intuitive visual effect for user.

1. Graph data

In the respect of variation trend of data, graph is more intuitive, accurate and effective for data display. TeeChart plug-ins can finish graphical demand in VC++, style of interface is more beautiful, however, from the perspective of simple and stable needs, system selects MSCHART graphics control of visual graphics statistics and reports, which has more simple operation. It selects the display type of image with SetChartType way such as bar graph (contrast diagram), pie graph (scale map) and curve graph, and so forth.

Due to the continuous characteristics of logging data [2], the data in the display possesses a certain demand on the video memory, for instance, MSCHART control outputs data at the same time, whose number is 12 dimension multiply 10 thousand, the control can not display, but the process exists. Dynamic display method just can solve this problem, which has the hardware environment of a 256 MB video card and 1 GB memory. Dynamic function of display image data can be completed by the OnTimer function. The temporary ID of data attribute in the table can be associated with sample counter in each sample, which can call the data required.

2. Image display and format conversion

At present, in order to facilitate network transmission, the logging images usually use compressed format of JPG and GIF file format, what’s more, BMP format, it is also irreplaceable, for its outstanding characteristic in Windows platform supporting and lossless compression, so the system uses the BMP, GIF, JPEG format for logging image processing.

From the image effect on accuracy, using lossless compression BMP format and MSCHART plug-in technology complete the task of formation, preservation and display of loading image for logging image. Fig. 3(a) is the dynamic graph of a certain well Su in BMP format in the 3200-3500 interval, whose attributes are GR, DT, WQ, after a series of operation by suspended, save and then generate, and display number of interface sets to 12, and system reads it in 50%.

In addition, there is format conversion module as well, for facilitating two other formats of image to read. The choice of system is calling GDI+ class library to finish it. For example, as shown in Fig. 3(b), the system reads, a certain well Xia in Xinjiang, logging graph of BMP format converted by GIF format in 50%.

5 The Practical Application Example

In order to verify the function and effect of the intelligent data mining system, we have actually processed a certain well Xia in Xinjiang to get reservoir identification. The depth is between 1220m and 1290m, it has oil. There are 10 attributes, i.e., AC, GR, NG, RA2, RA4, RI, RM, RT, RXO, SP. We extract sample point after normalizationin, adopt uniform sampling method, extract a depth point data from every two depth points, and contain the key breakpoint of reservoir. And then we do attribute reduction after discretization, the reduction results are AC, NG, R1, SP. We use SVM to establish reservoir recognition model, and use original sample data to identify whether
this well section has oil, Fig. 4(a) is classification diagram of Matlab combined programming, blue dot on behalf of the reservoir, a red dot is dry layer; Fig. 4(b) is the effect statistics and interpretation chart of this interval, the testing accuracy reaches 83.66%, consumes 41.61 seconds (hardware conditions: CPU frequency is 1.6GHz), depth between 1220m and 1290m is reservoir, its conclusion accords with conclusion of oil test. Apparently, this effect of data mining system is good.

6 Conclusion

This article has been developed a set of petroleum logging intelligent data mining system for the first time based on the VC++ and intelligent data mining technology. Utilizing the SVM modeling method completes the mining work of identification in a certain well Xia in Xinjiang. It possesses the functions of data management, graph data processing, image access and format conversion in scientific computing visualization.

The experimental results show that the developed intelligent data mining system possesses friendly interface, high visualization degree, strongly expansibility and good mining effect. Admittedly, this system still needs to be further improved and expanded.
Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 60972106) and Tianjin Natural Science Foundation (No. 11JCYBJC00900).

References


