Application and Research on Data Mining Technology in the Hospital Information System

Yang Hui
New Star Application Technology Institute
He fei, China
Lf2000_www@yahoo.com.cn

Wang Mei, Xue Xiangfeng, Chen Peibin, Dong Yan
New Star Application Technology Institute Hefei, China

***Abstract* — Most of the current medical database in hospital information system is poorly applied, the low level data integration and superficial analysis can not satisfy the automatic access to medical knowledge, not to mention the requirement of hospital management. This paper introduces the key technologies in medical data mining such as data pretreatment, anonymization, identity transformation, etc. It elaborates the basic process of data mining in hospital information system, including data extraction and pretreatment, mining algorithm execute, model discovering, knowledge expression and evaluation. It also concretely discusses the application of data mining technology in hospital information system in two major aspects: modal construction and implementation process.**

Keywords- Hospital information system(HIS); data mining; model; process

I. KEY TECHNOLOGIES IN HOSPITAL DATA MINING

A. Data pretreatment

Data pretreatment is an important step in the process of data mining (DM), especially when the database contains noisy, incomplete and in-consistent data. It usually takes about 60% time to treat the original data in the whole procedure of data mining, and only 10% the real time for key DM process. The main steps of data pretreatment include data cleaning, data integration, data transformation and data elimination.

B. Data anonymity and identity transformation

Data pretreatment is an important step in the process of data mining, especially when the database contains noisy, incomplete and in-consistent data. It takes about 60% time to treat the original data in the whole process data mining, and the time for data mining is only 10%. Data pretreatment includes mainly data cleaning, data integration, data transformation and data elimination.

Because medical information involves the problem of patients' privacy, the medical information needs special treatment, such as the anonymity of the patient and the identity transformation. Anonymity means to remove the real identity of the patient, or to replace the true identity with wrong identity. After the process of anonymity, researchers can not query any private information of patient through clinic records. Identity transformation differs lightly from anonymity. The identity after the transformation may still contains some real information of patient, and can be viewed only by the authorized researcher.

C. Data mining on medical text data

For the medical text information, it's hard for people to interpret the data, such as image, signal and other clinic data with non-standardized method, and difficulty to implement the data mining with these non-standardized data. At present, medical text data standard transformation works well with computer technology. The main three steps are: analysis of source statement, transformation and target statement creation. The main difficulty of transformation is that the source statement is not unique, so it needs to collect all types of source statements. The current machine transformation can treat sentence of less than 10 words. XML (Extensible Markup language), a structured language can be another way of text data standardization. XML can not only create structured text data, but also a good tool to treat and share data. XML is a key technology for data mining and knowledge finding.

D. Data mining on image data

The current medical images come from imagery machines, such Computed Tomography (CT) and type B ultrasonic, which have been proved a reliable assistant diagnose mean for the doctor. Different from data mining using pure digital data, it's more difficult to accomplish the data mining through medical images, so the development of effective image data mining tool becomes a key technology. Data mining through medical image includes the following aspects:

- Removal or reduction the influence of the image noise to enhance the target image quality and to fetch the border of the target organization.
- Description and characterization of the target organization on concept level, to acquire and validate the dynamic scope of some correspondent parameters;
- Management and index of the medical image data.

Nowadays, breakthrough of data mining to **SPECT** image has been achieved. Furthermore, research on rapid and high quality mining algorithm, to guarantee the accuracy and reliability of the knowledge through data mining are also the key factors of medical data mining.

II. THE PROCEDURE OF DATA MINING IN HIS SYSTEM

A. *Basic procedure of DM in HIS*

Figure1 shows the basic procedure of data mining in HIS.

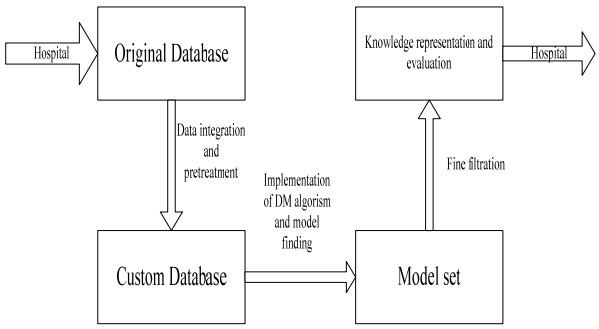

Figure.1 The basic procedure of data mining in HIS.

B. *Extraction and pretreatment of HIS original data*

It needs to extract correspondent data from different kinds HIS databases, files and non-electronic data source after the task of data mining has been confined according to decision made by the top administration of the hospital. The extraction can be achieved using data base query language. The extracted data then needs to be refined and pretreated because these data have been affected by noisy data, blank data and non- consistent data. The pretreatment procedure includes elimination and correction of wrong data, unification of data dimension and the extraction of time serial information, and so on. The last step of pretreatment is to transform and compress the data, such as to transform the formation of data expression, to extract the character of data and to decrease the demission of data. This step can supply the next model finding step with high quality input data.

C. *Excuting of the DM algrithism*

Due to the different application aims, there are different corresponding data mining task. For different mining task, the mentioned process can produce several different custom-built databases with corresponding data mining algorithm in the process of data mining. All the data mining algorithms are characterized by their own statistics parameters used as the evaluation standard, such as reliable degree, interest degree and novelty degree. The evaluation standard then is used to decide which models should be reserved or be discarded so as to find the potential best model.

D. *Model finding*

Model finding is the core stage of HIS data mining. In this stage, the overall models and rules will be found through analyses of historic data using different data mining algorithms.

E. *Knowledge Expression and evaluation*

Model finding is the core stage of HIS data mining. In this stage, the overall models and regulations will be found through analyses of historic data using different data mining algorithms. The final models and rules will be displayed by visualized and easy-understood media. At the same time, the models and rules should be evaluated and chosen with best performance.

III. APPLICATION OF DM TECHNOLOGY IN HIS

A. *Establishment of DM model*

DM model can be established using various modeling methods, such as the model guide of Analysis Services (AS), DSO and other software that can create AS or client data model. The dialog level DM model will be established with the same performance of theoretical model. The structure of local DM model is similar to the table in the relational database (RDB). Likes the table, DM model defines the DM content using “Column”. Differing from the table in SQL Server 2005, the column of model can be embedded with table. The Analysis Services of SQL Server 2005 support two kinds of DM models: cube model based on OLAP and model based on cross relation table.

1) *Model based on OLAP cube*

This paper created an OLAP DM model using code “CREATE OLAP MINING MODEL”, the code is defined as the following,

CREATE OLAP MINING MODEL <Model Name> From <Case Cube Name> (<Cube Members >) Using <Algorithm Name>

<Model Name> defines the model name. The physical location of model is defined by property of Mining Location. If the Mining Location property is not classified in the connecting string, the mining model created by the connecting string will be live only in the session of dialog. <Case Cube Name> is the name of testing case cube which contains the < Cube Members > model. The <Algorithm Name> shows the algorithm name by which the model created.

The following example creates an OLAP mining model, MyOlapModel, for medical diagnosis:

```
CREATE OLAP MINING MODEL〔MyOlapModel〕
FROM 〔Cure〕
(
CASE
DIMENSION 〔userID〕/*Patient ID*/
DIMENSION 〔userSEX〕/*Patient SEX*/
DIMENSION 〔userAGE〕/*Patient AGE*/
LEVEL 〔userNAME〕/*Disease NAME*/
PROPERTY 〔result〕/* diagnosis description*/
PROPERTY〔inhospitaldate〕PREDICT /*patient visiting date time*/
)
USING Microsoft – Decision – Trees
```

2) *Model establishment with tables of RDB*

The relational mining model is defined by Column within the designed model based on the tables in RDB. For the unknown data format and structure, every Column is defined by name, data type, statistics character and the query predictability. The common procedure to create relational mining model can be expressed with the following code,

CREATE MINING MODEL <Model Name> (<Column Members>) USING <Algorithm Name>

For instance, the following mining model is defined:

CREATE MINING MODEL 〔Member Cards〕
(
〔userID〕 LONG KEY,
〔userAGE〕 INT,
〔inhospitaldate〕 DATE,
〔result〕 TEXT DISCRETE,PREDICT
)
USING Microsoft – Decision – Trees

In this example, the CREATE MINING MODEL sentence is used to describe a mining model named as Member Cards. The syntax is similar to the CREATE TABLE sentence used in SQL. The columns are defined and classified by their own data content and extra information. The Result column is specified as predictable by using PREDICT keyword.

B. DM implement in the client terminal

1) Procedure summarization

The rich data mining functions can be accessed with thin-client application created by JAVA API. Complied with JSR-73, the standard extension frame, ODM Java API extends SQL to the special JDM1.0 which is the Java API industrial standard established by JCP (Java Community Process). JDM1.0 defines the JAVA interface for the engine of data mining. The mining functions supported by JDM interface include classification, regression analysis, clustering, and property value and correlation analysis. The mining algorithms include naive Bayesian classification, supporting vector machine, decision tree and K mean value. In this paper, the implement procedure of data mining technology is consisted of the following steps: data collection, data pretreatment, engine execute and treatment of result, as shown in figure2.

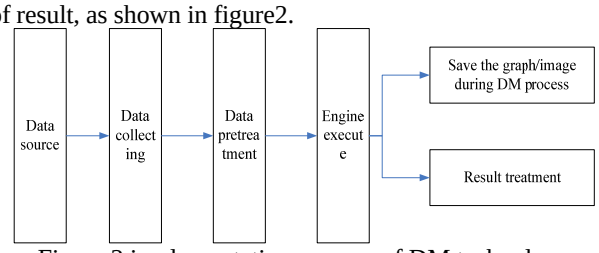

Figure.2 implementation process of DM technology

2) Data collection

The user submitted files are received by the way of uploading. User data files like Excel file format can be uploaded through graphic user interface, and then be analyzed by the specific program. The text code in the first line of file acts as the input/output variable.

3) Data pretreatment

In this step, the data from different sources will be filtered and then be integrated through data cleaning and data transformation. There exist similar and repeated data from different data source, and the same property may be expressed in different way, so data cleaning should solve the problem of property redundancies and value collision. The steps include: recognizing the repeat property identifying the same character; dominating of approximately repeated data and deleting extra property from dataset. Data with different expression method and different precision from heterogeneous databases need to be formatted. After the data transformation, the new value will be stored in sample data series.

4) Eninge execute

It is the core step of DM process. The key aspects include control of input and output data, the selection of analysis method and building of the DM script code. The executing program will create in real-time an execution script and then transfer it to DM software (such as SPSS Clementine), so the analysis result of every step can be made with this script code. The script content includes data source, input and output parameter, analysis model and output result.

During the process of DM analysis, there are several types of node with different implement method and characteristic, so it needs to abstract the common character of node and the public interface of analysis algorithm. When the user operates on the interface, he can combine the different selected nodes, and the individual parameter of some node can also be set. In the end, the execute code will be generated by extract procedure, as shown in figure3.

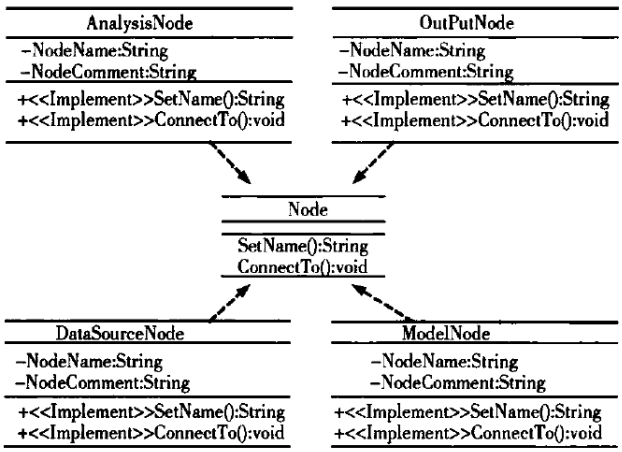

Figure3. UML of DM script

Executing engine can be supported by SPSS Clementine software. When the executing engine starts to run the calculation, one management thread from thread pool (Thread P001) will be started to set the new analysis request object, and to start a new analysis process of SPSS Clementine mining module. Complex calculation and analysis will be implemented by specific process and be monitored by management thread. After the completion of analysis process, the management thread will take back and store the result into database, and inform the user when some errors have been made. In the end of analysis process, the management thread will delete the analysis request object and send back to the thread pool, as shown in figure4.

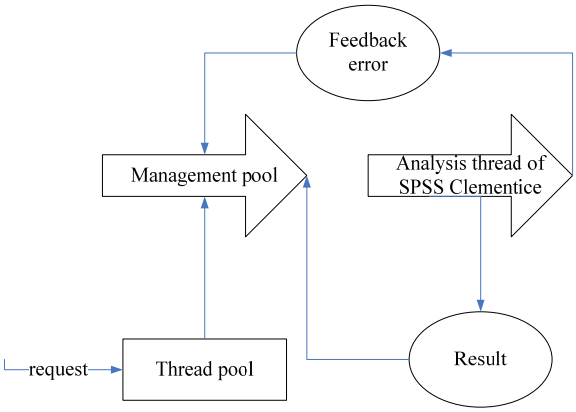

Figure4. Calling relationship between executing engine and DM software

The last step of DM is the treatment of DM result, the graph and tables will be stored in the database.

IV. CONCLUSION

DM technology combined with machine learning, statistic analysis and database theory, can find the model under the data of HIS database, and helps the manger to enhance the HIS data utility efficiency by analyzing the large amount of HIS data with DM model, and to improve the serving quality of hospital, and to enhance the decision performance of manager. In this paper, the key technologies in medical data mining such as data pretreatment, anonymization, identity transformation, etc. has been introduce, and the basic process of data mining in hospital information system, including data extraction and pretreatment, mining algorithm execute, model discovering, knowledge expression and evaluation are elaborated and discussed. The two key aspects, including modal construction and implementation process have also been discussed and shown in this paper.

REFERENCES

[1] Liang Xun, Data Mining algorithm and application, China: Bejing university press, 2006

[2] Zhao Xiao-fan, Niu Cheng-zhi, “Research on the Application of Data Mining in HIS Based on Decision Tree” Computer Knowledge and Technology, Hefei,China, Vol.7, No.2, January 2011,pp.292-294

[3] Wu Deyi, “The Application and role of Data Mining technology in hospital management” Chinese Medical Machine Information, Bejing,China, Vol.15, No.7, January 1999,pp.67-72

[4] Zhang Ning, Chen Yang, “The Application Data Mining technology in hospital management” Hospital Management Forum, Bejing,China, Vol.10, No.28, January 2011,pp.55-57

[5] XU Yuan-xi ZHANG Jie, “Application and Research of Data Mining in the Hospital Information System”, Micro Computer Information, Bejing,China, Vol.14, No.11, January 2008,pp.188-190

[6] Hou Jie, Zhang Xi-kun, “Research on the Application of Data Mining in HIS Based on Decision Tree” Computer Knowledge and Technology, Hefei,China, Vol.7, No.30, October 2011,pp.7365-7366

[7] Zhu L ingyun, Wu Baom ing, Cao Changxiu, “Research on the technology,algorithm and application of medical Data Mining” J Biomed Eng Bejing,China, Vol.20, No.3, October 2003,pp.559-562