Design of Medical Examination Data Mining System Based on Decision Tree Model

Gang Ma1,*, Liumei Zhang1, Guangzhi Cui2, Yimin Cheng2
1School of Computer Science, Xi’an Shiyou University, Xi’an, Shaanxi, China.
2Hospital of Xi’an Shiyou University, Xi’an, Shaanxi, China.
*jhimg@xsyu.edu.cn

Abstract. This paper analyzes the system of OLAP physical examination data mining system, the concept of classification, and the nature of decision tree to define the implementation process of ID3 algorithm. Transaction data of the physical examination information management system of Xi’an Shiyou University Hospital from 2015 to 2018 are selected and prepared for cleaning and preprocessing, then the data is dimensionally reduced and transformed, to design a direct data model suitable for classification analysis. The sampling data set is classified and analyzed by ID3 algorithm, and the classification rules are extracted. Using the prediction conclusions of these classification rules, the medical examiner can quickly and scientifically predict the individual physique of each college teacher, so as to classify college teachers of different physiques, and specifically guide the teacher's further treatment plan of exercise, diet and work. It has high reference value and promotion towards teachers' health.

1. Introduction
The traditional medical examination management information system based on ON-Line Transaction Processing (OLTP) is oriented to the specific application requirements of the daily business of the medical examination organization. The design is mainly concerned with the performance of transaction processing (mainly represented friendliness and response time of the physical inspection business processing operation) to ensure the speed of data acquisition, the correctness, security and efficiency of data processing, and optimize the efficiency of data storage, query and update[1]. The main task of OLTP emphasizes updating the database, which is to add records to the database.

The on-line Analytical Processing (OLAP) based medical examination data mining system is specially designed for specific problems. It is used to support online data access and analysis of complex analysis operations. It is a verification-type analysis software system with inductive function. It focuses on decision-making support for decision-makers and senior management. It can respond to analysts' requirements for complex query processing of large data volumes quickly and flexibly, and provide query results to decision-makers in an intuitive and understandable form to achieve goal of enhance accuracy of the operation status of medical examination institutions, understand the needs of medical examiners, formulate correct programs, and increase benefits [2].

The OLAP-based medical examination decision support system mainly realizes the scientific and efficient information management of medical examination institutions, by which the data mining tasks can be conducted on the accumulated massive data to discovers knowledge, providing accurate decision-making information for decision-makers and medical examiners. OLTP medical examination
information management system is fundamental to providing raw data for the OLAP physical examination data mining system; on the other hand, the OLAP physical examination data mining system provides a broad space for the expansion of the OLTP medical examination information management system. The organic combination of the two systems provides effective technical support for the business management and data mining of the medical examination organization. It is indispensable to improving the efficiency, management level and economic efficiency of the medical examination institutions [3].

2. Classification analysis
Classification analysis is a very important task of data mining, which reflects the characteristics of the commonalities and differences between different things. Classification is mainly used to predict features to elicit set of models or functions that can describe the typical features of data sets, so as to identify the attribution or category of unknown data.

Classification model can be learned from a set of training sample data by data mining classification algorithm. The construction methods of classification model include decision tree method based on machine learning, Bayesian method based on statistical principle and non-parametric method, rough set algorithm based on rough set theory and BP algorithm based on neural network. This paper mainly uses ID3 algorithm of decision tree classification analysis to mining daily physical examination business data of physical examination institutions, finds out the classification rules of personnel, and forecasts the health status of different types, so as to guide the diet, rest, exercise and treatment plan of the personnel.

2.1. Decision Tree
Decision Tree is a directed acyclic graph (DAG)[4]. Nodes without input edges in the tree are called root nodes. Nodes with input edges and output edges are called internal nodes. Nodes with input edges and without output edges are called leaf nodes. Decision tree learning is a case-based inductive learning method, which infers from a group of disordered and irregular cases and constructs classification rules formed by the representation of decision tree. The top-down recursive method is used to compare the attribute values of the internal nodes of the decision tree and to judge the branches from the node down according to the different attribute values. The conclusion is drawn at the leaf nodes of the decision tree [5].

For the original training tuple data set, which classification order, i.e. which attributes are used as classification criteria in turn, makes the final decision tree most beneficial to the generation of classification rules. It is the most critical problem in the process of constructing decision tree. The attribute selection order is called attribute selection measure. In this paper, the information gain, gain rate and Gini parameter are compared to construct decision tree by choosing different indicators under different conditions. Because all the attributes involved in the decision tree classification of medical examination data are discrete, the following discussion only focuses on discrete values.

2.2. ID3 algorithm
The ID3 algorithm adopts a greedy non-backtracking method, and the decision tree is constructed in a top-down recursive division. Classification begins with training tuple sets and their associated class labels. As the tree is constructed, the training set is recursively divided into smaller subsets. The core of the algorithm is to select attributes on all levels of nodes by using the highest information gain, maximum information gain rate and minimum Gini parameters as the criteria. When testing each non-leaf node, the maximum class information about the tested example can be obtained, and the entropy value of the system is the minimized when the sample set is divided into subsets by using this attribute. It is expected that the average path of the non-leaf node to each offspring leaf node is the shortest, so that the average depth of the generated decision tree is smaller, and the classification speed and accuracy are improved[6-7].
3. System analysis and design
The OLAP-based medical decision support system starts with the existing data source of the transaction of the traditional OLTP based physical examination management information system. It carries out a series of data processing and forms corresponding models. These models are presented and interpreted then ultimately translated into knowledge that is understandable to the observers for management decisions in medical institutions. The process is an interactive and iterative data mining process that requires a large number of decision-making interventions by analysts.

The process model of provides a macro guidance and engineering method for data mining then enabling decision analysts to better research, develop and use knowledge. Brachman et al. proposed a more practical process view in 1996, which was further developed and improved by Fayyad et al., and proposed the Fayyad process model and was widely accepted because of its versatility[8]. As shown in Fig. 1, the process model of physical examination data mining and online analysis processing identifies the knowledge represented by rules from the medical examination data set, which is an advanced data processing process. It consists of the following six steps, each of which interacts with each other and may have multiple iterations, so it is a spiral ascending process as a whole.

3.1. Define the subject
The ID3 algorithm is used to construct the decision tree, and the relationship of each physical examination item in the dataset is mined, then the classification rules between the item data and the health status of the medical examination personnel are determined. That is, the rule is elicited as "IF systolic blood pressure = normal AND diastolic blood pressure = normal AND blood glucose = normal AND cholesterol = normal AND triglyceride = elevated AND high density lipoprotein = normal AND low density lipoprotein = normal Then health assessment = high glycerol Classification rules for triglyceridemia. Applying these rules, the medical examiner can pre-judges the health status according to some important physical examination items of the medical examiner, and support them further treatment plans and lifestyle habits such as exercise and diet. According to this research question, the identified themes are: personnel, items, item indicators, medical examination data and doctors.

3.2. data preparation
Based on the identified topics, following data need to be prepared and collected from the transaction database:
Medical examiner information (health card number, name, gender, birthday, ethnicity, birthplace, marriage, type of person, ...), medical examination information (serial number, name, ...), medical
3.3. Data selection
According to the goal of mining, the datasets related mining task are selected from the physical examination transaction database and then integrated to form the target data. It also solves the problem of data format differences caused by data types, operating systems, and data mining platforms.

In order to prepare and collect subject-related data from the transaction database we need to extract the attribute fields as follows: health card number, name, gender, birthday, ethnicity, birthplace, marriage, crowd type, department, Group item, medical examination item, physical examination status, physical examination result, examination prompt, time, reference value type, reference value unit, reference value upper limit, reference value lower limit, high limit time complications, bottom time complications, project evaluation, evaluation description, evaluation Suggestions and so on.

3.4. Data cleaning and preprocessing
The incomplete, inconsistent, inaccurate and redundant "dirty data" in the target data are analyzed by the classification rule-based method. The corresponding cleaning and preprocessing operations are implemented to form a purified data set for ID3 based algorithm.

Data cleaning and preprocessing mainly include data verification and data mapping. The function of data verification is to remove the records that do not meet the test conditions before entering the data warehouse. The physical examination status field identifies the business flow of the entire physical examination from registration to review complete. These states are: registered medical examination, in physical examination, complete physical examination, review medical examination and history examination. Among those data, only 'Complete Medical Examination', 'Review Medical Examination' and the 'History Physical Examination' are suitable for mining, all other states should be eliminated. The function of data mapping is to unify all data attributes (such as field name, data type and field width) before entering the data warehouse. Since the author has solved this problem in the design of the transaction database, it is not considered here.

3.5. Data cleaning and preprocessing
The data in the transaction database is stored according to the application requirements of the daily physical examination and management business of the medical examination institution. The design and optimization of the structure focus on the efficiency of transaction processing such as response
time. The data of each topic in the data warehouse is stored according to the requirements of the decision analysis application. The design and optimization are more concerned with the integrity of the data, the consistency, and the accuracy of the analysis results. Therefore, before loading the data into the data warehouse, it is necessary to perform the aggregation and transformation operations to meet the needs of data mining. The logical structure is shown in Figure 2.

3.6. Data mining

Data mining is to select appropriate techniques and algorithms for data analysis. It has two tasks including determining the function of data mining and selecting the appropriate pattern search algorithm. This paper uses the ID3 algorithm of classification analysis function to realize data mining of physical examination data.

The mining task use the data from the transaction database of the physical examination information management system of Xi'an Shiyou University Hospital. Of all the data record, 5000 adult male medical records from 2015 to 2018 are selected and forms statistic data set R (limited to the length, only partial sampling data is given here), see Table 1. Then data set R is sorted and cleaned, for conduct index conversion of data of all medical examination items to further extract training set R’ according to medical examination item indicators, as shown in Table 2. As can be seen from the description of the ID3 algorithm, the core idea is the selection measure of the attribute. The author uses the Health Evaluation (HE) as the class label attribute, and selects the Systolic Blood Pressure (SBP) and the Diastolic Blood Pressure (DBP), Blood Glucose (GLU), Cholesterol (CHOL), Triglyceride (TRIG), High-Density Lipoprotein (HDL), Low Density Lipoprotein (Low-Density) Lipoprotein(LDL) and other 7 attributes related to the three high risk diseases (hypertension, hyperlipidemia and hyperglycemia). To achieve the goal of classification of the physical examination data of all working male teachers of Xi'an Shiyou University.

| ID  | SBP (mmHg) | DBP (mmHg) | GLU (mmol/L) | CHOL (mmol/L) | TRIG (mmol/L) | HDL (mmol/L) | LDL (mmol/L) | HE                     |
|-----|------------|------------|--------------|---------------|--------------|--------------|--------------|------------------------|
| P0001 | 197        | 116        | 5.42         | 4.21          | 2.21         | 0.97         | 2.80         | Grade III hypertension |
| P0001 | 116        | 67         | 5.24         | 4.80          | 2.41         | 1.09         | 2.93         | Hypertriglyceridemia   |

| ID  | SBP (mmHg) | DBP (mmHg) | GLU (mmol/L) | CHOL (mmol/L) | TRIG (mmol/L) | HDL (mmol/L) | LDL (mmol/L) | HE                     |
|-----|------------|------------|--------------|---------------|--------------|--------------|--------------|------------------------|
| P0001 | Severely elevated | Severely elevated | normal        | normal        | normal        | normal       | normal       | Grade III hypertension |
| P0002 | normal     | normal     | normal        | Rise          | normal       | normal       | Hypertriglyceridemia  |

It is the premise of on-line analysis and processing to evaluate the patterns discovered by data mining according to the decision-making needs of users and to display useful patterns and their data to users in an effective form, and to evaluate and interpret the validity of patterns and their results. In this paper, the classification rules are extracted as follows:

**Rule 1:** IF SBP=Severely elevated AND DBP=Severely elevated AND GLU=normal AND CHOL=normal AND TRIG=normal AND HDL=normal AND LDL=normal Then Health assessment= Grade III hypertension

**Rule 2:** IF SBP=normal AND DBP=normal AND GLU=normal AND CHOL=normal AND TRIG=rise AND HDL=normal AND LDL=normal Then Health assessment= Hypertriglyceridemia
4. Conclusion
In this paper, data selection, data preparation, data cleaning and pre-processing, data dimension reduction and conversion are carried out for the transaction database of the medical examination information management system of Xi'an Shiyou University Hospital from 2015 to 2018. Then a direct data model suitable for classification and analysis is designed. ID3 algorithm, a typical algorithm of classification analysis, is used to classify and analyze the sampled data sets, and the classification rules are extracted. Through comparison, it is found that these classification rules are highly consistent with the clinical practice of doctors, and highly consistent with the individual survey results of teachers. Therefore, using the prediction conclusion of these classification rules, the physical examiner can quickly and scientifically predict the individual physique of each college teacher, so as to classify the college teachers with different physiques, and guide the teachers' further treatment programs and living habits such as exercise, diet and work and rest. The method has very high reference value and promoting role.

Acknowledgments
This work was supported in part by the scholarship from China Scholarship Council(CSC) under the Grant CSC No.201809910007.

References
[1] SU Xin-ning, YANG Jian-lin, et al. Data Warehouse and Data Mining[M]. Tsinghua University Press, Beijing, China, 2006.
[2] XU Jie-pan. Data Warehouse and Decision Support System[M]. Science Press, Beijing, China, 2005.
[3] MA Gang. Application of Data Mining Technology in Hospital Management Information System[D]. Xi'an Shiyou University, 2008.
[4] Johannes Gehrke, Venkatesh Ganti, Raghu Ramakrishnan and Wei-Yin Loh. “BOAT: Optimistic Decision Tree Construction”. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, Philadelphia, Pennsylvania, May 1999.
[5] MA Gang, LIU Tian-shi, LI Jiao. Application study of ID3 algorithm in mining of doctor classification regulations[J]. China Academic Journal Electronic Publishing House, 2008 (11) 79-81.
[6] Zhang Quancheng, You Kun, Ma, Gang. Application of ID3 Algorithm in Exercise Prescription[C]. The International Conference on Electric and Electronics, Nanchang, China, June 22, 2011. 99(3): 669-675.
[7] LI Shoubang. Application Study on Mining of University Students’ Physical Fitness Test Data Based on Classification Rules: Taking the Juniors of Xi'an Shiyou University as an Example[J]. Journal of Xi'an Shiyou University(Natural Science Edition), 2018, 33(5) : 120-126.
[8] Fayyad U M, Platetsky Shapiro G, Smyth P, Uthurusamy R U. Advances in Knowledge Discovery and Data Mining[M].[S.1.]:AAAI, 1996:83-115.