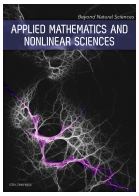
sciendo Applied Mathematics and Nonlinear Sciences 5(2) (2020) 443–452



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

# Deep Integration of Health Information Service System and Data Mining Analysis Technology

Zhihao Cui[1,2], Chaobing Yan[2,3†]

[1]Pingdingshan Universtiy, Dept Sports Sci, Pingdingshan, Henan, China

[2]Wonkwang University, Dept Sports Sci, Iksan Si, South Korea

[3] Jiujiang University, Dept Sports Sci, Jiujiang, Jiangxi, China

**Submission Info**

Communicated by Juan Luis García Guirao
Received June 4th 2020
Accepted August 14th 2020
Available online December 21st 2020

## Abstract

The scale and complexity of health information service system has increased dramatically, and its development activities and management are difficult to control. In the field of, Traditional methods and simple mathematical statistics methods are difficult to solve the problems caused by the explosive growth of data and information, which will adversely affect health information service system management finally. So, it is particularly important to find valuable information from the source code, design documents and collected software datasets and to guide the development and maintenance of software engineering. Therefore, some experts and scholars want to use mature data mining technologies to study the large amount of data generated in software engineering projects (commonly referred to as software knowledge base), and further explore the potential and valuable information inherently hidden behind the software data. This article initially gives a brief overview of the relevant knowledge of data mining technology and computer software technology, using decision tree graph mining algorithm to mine the function adjustment graph of the software system definition class, and then source code annotations are added to the relevant calling relationships. Data mining technology and computer software technology are deeply integrated, and the decision tree algorithm in data mining is used to mine the knowledge base of computer software. Potential defect changes are listed as key maintenance objects. The historical versions of source code change files with defects are found dynamically and corrected in time, to avoid the increase of maintenance cost in the future.

**Keywords:** health information service system, computer software, data mining, decision tree, data analysis
**AMS 2010 codes:** 05-04, 68P20, F18423

## 1 Introduction

It is one of the important topics for future research to innovate medical and health services by using the development of computer software technology and information technology. The developed countries like the

†Corresponding author.
Email address: yan8085@126.com

sciendo
ISSN 2444-8656 doi:10.2478/AMNS.2020.2.00063

United States have been studying this for decades. On April 27, 2004, president Bush of the United States issued Presidential Decree 13335, which explicitly stated that all the health records of US should be digitised within 10 years which means all the records will become electronic health records (EHRs). In 2005, the United States Government established an advisory committee, AHIC, to provide recommendations and solutions to the problem of medical informatisation in the United States. In 2009, President Barack Obama promulgated President's Decree 13507, promulgating the HITECH Act as part of the health care reform of the United States; in 2010, the United States signed the Patient Protection and Affordable Medical Act again, followed by the creation of the Federal Medicare and Medicaid EHR incentive programs. Thus, the United States pays more attention to the application of electronic health records and electronic medical records in when shaping the health information policy. Japan is the leading nation in Asia in terms of health care policy and informatisation. Japan's National Health Insurance in 1961 and the Elderly Health Act in 1983 made public social insurance available to all citizens. The European Commission's 'Action Plan for European Informatization 2002 and 2005' calls for a focus on medical informatisation, which is strongly supported by the European Council. In May 2003, the European Commission and Member States put forward the concept of 'High Level Conference on Medical and Health Informatisation'. The purpose of these meetings is to provide expertise and knowledge to leaders and to show the latest technological achievements in the construction of medical informatisation in the European Union. In the same year, the Ministers of Health at the Brussels Conference announced their firm commitment to building information-based health care. Ireland introduced the European Action Plan on Health In 2004, and set the EU targets for 2008 and beyond. In 2017, the World Health Organization revised the Strategy for the Prevention and Control of New Diseases in the Asia-Pacific Region, proposing to improve the efficiency of cross-sectoral emergencies through health integration mechanism, systematic coordination and information sharing mechanism.

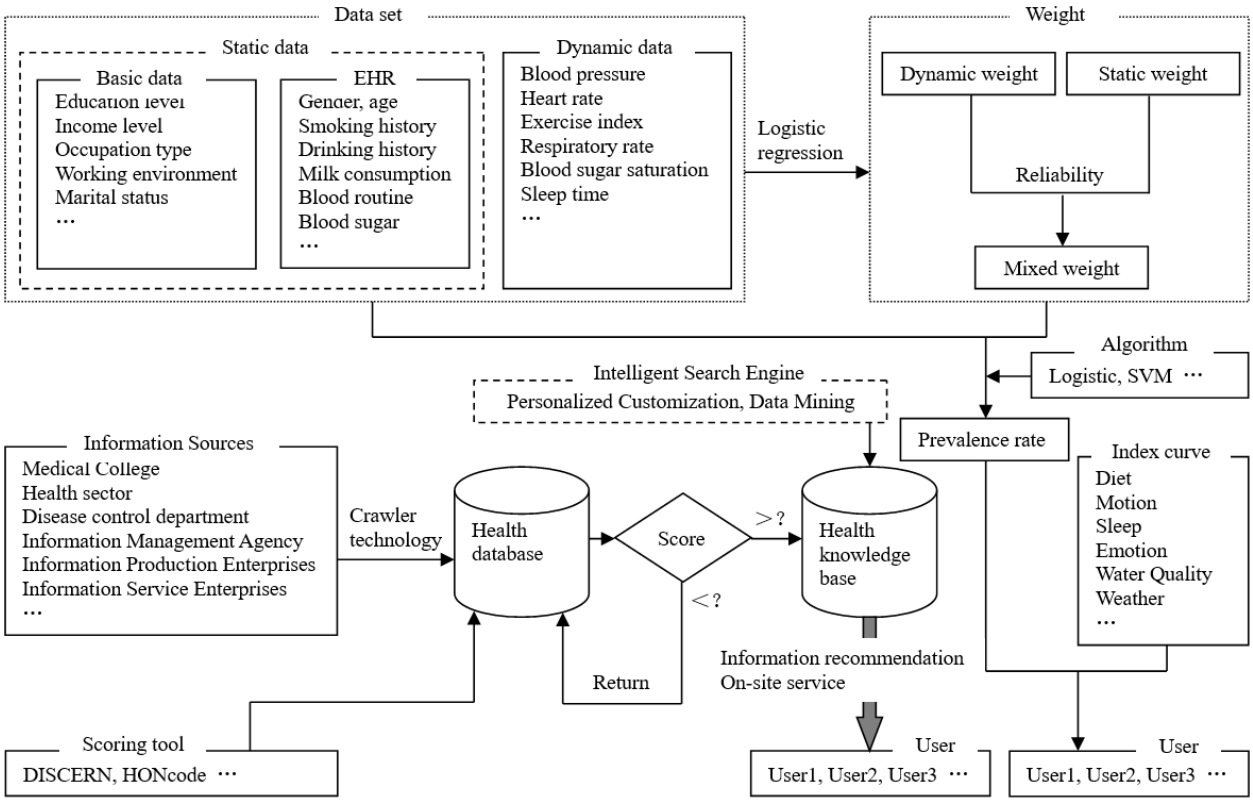



**Fig. 1** Technical Roadmap of Health Information Precision Service Model

Currently, the most common health information push tool is mobile terminal software. Health APP has gradually become the most important application of mobile phones with the rapid development of smart terminal devices. Users can simply check health data, consult health problems online, search health knowledge base, etc. Some personal health management software also has the functions of 'mass messaging' and 'intelligent reply' to ensure the authenticity and uniqueness of users. As shown in Figure 1, Technical Roadmap of Health Information Precision Service Model.

On the other hand, with the progress of society and the rapid advancement in science and technology, computer, communication and Internet technology have penetrated into all sectors and are changing everyday life of the mankind [1,2]. Health information service system generates, collects and stores a large amount of data using the new technologies in computer field. The ever increasing amount of data has become a perennial problem for various industries. A lot of information brings benefit to all sectors of society, but also brings a lot of problems.
(1) Information is too much to digest.
(2) It is difficult to identify the true or false information.
(3) Information security is difficult to guarantee
(4) Information forms are inconsistent and difficult to deal with uniformity.

Current database systems can efficiently execute data entry, query, statistics and other functions, but cannot find the causal relationship and rules in the data and cannot predict the future development trend based on the existing data. So, people began to put forward a new slogan: 'Learn to discard information'. At the same time, people began to think, 'How can we not be inundated with information, but find useful knowledge in time and improve the utilisation rate of information?' Faced with the challenge of 'abundant data and lack of knowledge', data mining technology has emerged at the historic moment, accompanied by the emergence of new computer technologies and new theories, and thriving in the information era. Its application in telecommunications, banking, biology, fraud, supermarkets and other fields shows its strong vitality [3–5].

## 2 Relevant Research Based on Computer Software Technology and Data Mining Technology

### 2.1 Characteristics of Computer Software Technology

Computer software technology refers to the program or related data set up in order to ensure the normal operation of the computer. Software is the interface between users and hardware. It is the core component of maintaining the normal operation of the computer. It is also the channel of communication between users and computers. It can improve the overall structure of the computer, meticulousness and reliability [6–8]. Computer software Technology is a type of computer technology, including data processing, artificial intelligence, process control and scientific calculation. It has the following characteristics:
(1) The cost of development continues to increase.
(2) The difficulty of development is increasing.
(3) The internal structure is becoming more and more complex.

### 2.2 Technical Analysis of Data Mining

Data Mining is the process of mining required data from large amounts of data stored in databases, data warehouses, or other storage information bases. It is generally accepted that data mining is a process of extracting hidden, potential, effective, novel, useful and ultimately understandable knowledge from many incomplete, noisy, fuzzy and random data [9, 10]. The general flow of data mining is shown in Figure 2.
(1) Data preprocessing

It includes raw data collection, data cleaning, data extraction and data transformation. The purpose of raw data collection is to determine the object of operation of discovery task, namely target data. It collects data according to specific needs and requirements.
(2) Data Mining

Data mining first determines the tasks of mining, such as data summary, classification, clustering, association rule discovery, sequential pattern discovery and so on. After the task is determined, it is necessary to decide the mining algorithm to be used. The same mining task can be implemented using different mining algorithms.

(3) Model Evaluation and Knowledge Representation

Patterns are the results of data mining. The interesting patterns that represent knowledge are identified based on the some measure of interest. This process is called pattern evaluation, in which the specific values of the measures are given by specific domain experts, users or domain standards. The result is 'knowledge', which can be submitted to users through visualisation technology or by converting the results into user-friendly representations.

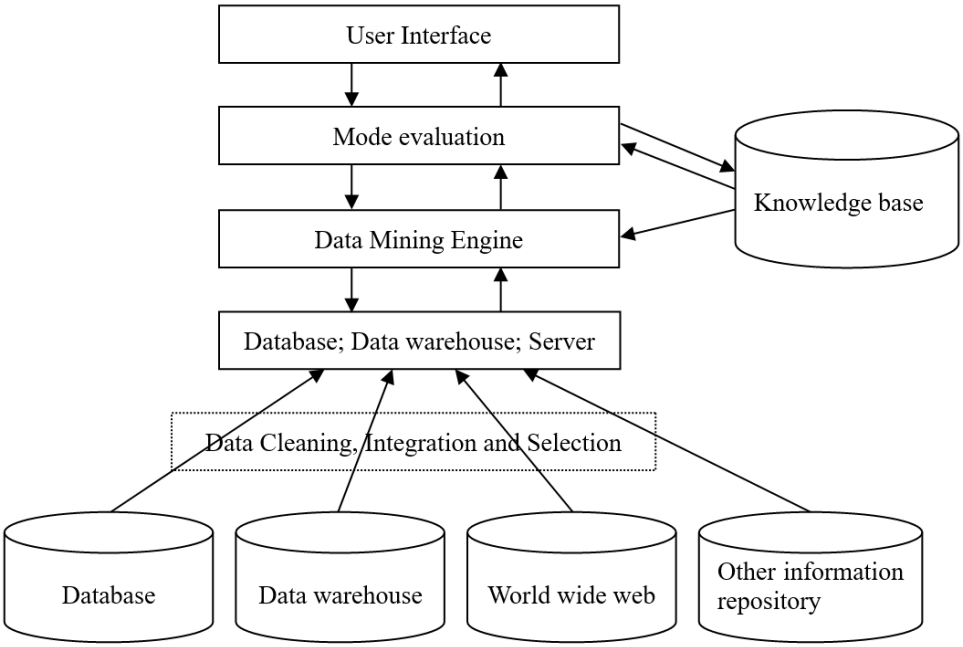



**Fig. 2** Data Mining Process

## 2.3 Construction of Health Information Resource Module

Currently, every country has been strengthening the construction of information resources services and standardising the content of information services continuously. Based on the research of other scholars, this paper builds an accurate health knowledge base, mainly from four aspects to ensure its accuracy. As shown in Figure 3, there is a multi-faceted and authoritative source of health information. We develop a set of web crawler technology to update the knowledge base in real time, select content scoring tools to measure the quality of information, and integrate heterogeneous digital health resources.

Health knowledge base has a wide range of information sources to extract more accurate and valuable health information from more comprehensive information. Faced with numerous health information sources, this study designed a web crawler program to obtain the latest health information released by various information sources in time. Current web crawler technologies include: General Purpose Web Crawler, Focused Web Crawler, Incremental Web Crawler and Deep Web Crawler. 'General Purpose Web Crawler' crawls a wide range, a large number, and is generally used to crawl search engines, but due to commercial reasons, it is not completely open to the public; 'Focused Web Crawler' is a technology applied on the premise of identifying crawling topics, so it is also known as 'Theme Network'. Because the purpose of crawling is very clear, it only crawls the

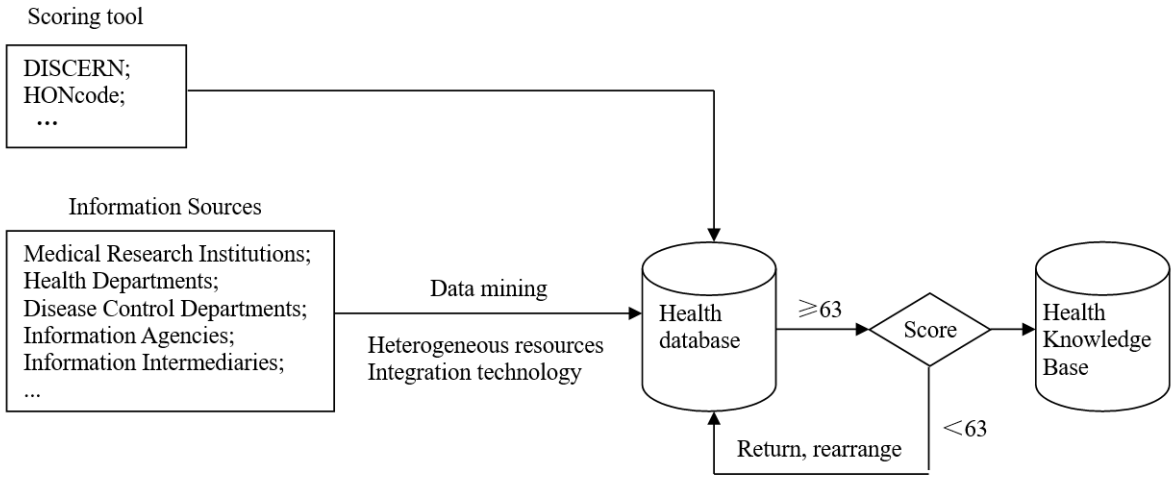



**Fig. 3** Process of Health Information Resource Module

web resources related to the theme, so it has fast crawling speed and less hardware occupation; 'Incremental Web Crawler' is to crawl new pages after determining the crawl, which is more real-time than periodic crawling; 'Deep Web Crawler' is a needle. For some deep web pages, such as those accessible only by using user accounts, there is a requirement for the user's right of access to the web pages. Most of the information sources of the health knowledge base constructed in this study are the websites of some government agencies and universities. So the amount of information released in a single day will not be much, the task of crawling is not very large, and the process control of crawling will not be cumbersome. Combining with the characteristics of this study, the strategy of 'Focused Web Crawler + Incremental Web Crawler' is selected to crawl health information sources.

As the part of data integration, this research integrates three technologies, XML, Web Service and Message Middleware, to integrate heterogeneous data. First, the heterogeneous data sources are shielded by message middleware technology. Then, the standard XML format data is generated by client software. Then Web Service is used for data integration. The processed data is output by middleware for storage and invocation.

## 3 Establishment of Data Mining Algorithmic Model

### 3.1 Overview of Decision Tree Algorithmic Model

Decision tree learning is one of the most widely used classification algorithms in the field of data mining. It is a method of approximating discrete-valued functions. It has good robustness to noise data and its rules are expressed in disjunctive expressions which are easy to understand. Since the 1960s, decision tree method has been widely used in classification, prediction, rule extraction and other fields. Especially after Quilan proposed ID3 algorithm, it has been further applied and developed in the field of machine learning and knowledge discovery. Further, the representative decision tree methods include CART, SLIQ and SPRINT. In this paper, we adopt simple ID3 algorithm. In ID3 algorithm, a significant improvement of CLS algorithm is that it determines 'certain rules of selecting test attributes' in CLS algorithm as selecting test attributes based on information gain. In ID3 algorithm, the window method is introduced for incremental learning to solve the problem that when the training instance set is too large, all data cannot be put into memory.

### 3.2 Algorithmic Model Computation

The flow chart of the algorithm is shown i n Figure 2. Let $T$ be a set of $t$ data samples. Its class attributes can take n different values, corresponding to $n$ different classes $C_i, i \in (1,2,3,\ldots,n)$. Assuming that $S_i$ is the

number of samples in category $C_i$, the amount of information needed to classify a given sample data object is:

$$I(t_1,t_2,\cdots,t_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

Among them, $p_i$ is the probability that any sample belongs to $C_i$, and it is advantageous to $S_i/S$ calculation.

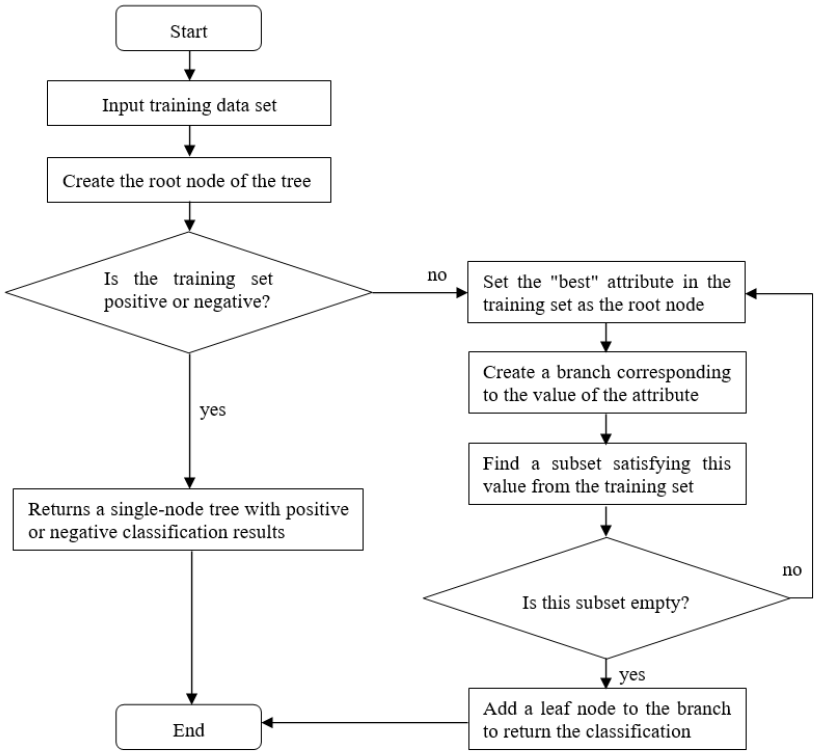



**Fig. 4** Algorithm flow

Let attribute $A$ have $k$ different values $(a_1,a_2,\ldots,a_k)$ and attribute $A$ divide $T$ into $k$ subsets $(S_1,S_2,\ldots,S_k)$, where samples contained in $T_j$ have a value $a_j(j=1,2,\ldots,k)$ on $A$, if attribute A is selected as the test attribute (i.e., the best partitioning attribute), then these subsets correspond to the branches growing from the set $S$. Let $T_{ij}$ be the sample number of class $C_i$ in subset $T_j$. It is known 'the smaller the entropy, the higher the purity of subset partition'. Entropy of subset partitioned according to $A$ or expected information is given by the following formula:

$$E(A) = \sum_{j=1}^{v} \frac{t_{1j}+\ldots+t_{mj}}{t}(t_{1j},\ldots,t_{mj}) \tag{2}$$

$$I(t_{1j},\ldots,t_{mj}) = -\sum_{i=1}^{n} p_{ij}\log_2(p_{ij}) \tag{3}$$

The encoding information to be obtained by branching on $A$ is:

$$Gain(A) = I(t_1,t2,\ldots,t_m) - E(A) \tag{4}$$

That is to say, *Gain* is considered to be the reduction of information entropy obtained by dividing the sample set according to the value of attribute $A$. ID3 algorithm calculates the information gain of each attribute and

**Table 1** Five Attributes and Interpretation of Bug Report Constructing Classifier

| Attribute types | Number | Attribute | Describe |
|---|---|---|---|
| Personnel attributes | $A_1$ | Solution personnel | Repair Bug Registrar |
| Bug Report and Solution Properties | $A_2$ | Bug priority | Priority level |
| | $A_3$ | Bug state | Current status of Bug reports |
| | $A_4$ | Modification instructions | Problem description |
| Code Modification Call Properties | $A_5$ | Notes | $F_1 \to F_3$<br>$F_1 \to F_3 \to F_4$<br>$F_1 \to F_2 \to F_4$<br>... |

**Table 2** Source Code Entities – Numbers corresponding to sequence of function operations

| Login () | Show () | Modify () | Statistic () | Add () |
|---|---|---|---|---|
| $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_5$ |

selects the attribute with the largest information gain as the test attribute of a given set $S$, and decides to generate the corresponding branch nodes. The generated nodes are marked as the corresponding attributes, and the corresponding branches are generated according to the different values of this attribute. Each branch represents a divided sample subset. The information gain of attributes describes the concept that for a training sample set $T$, the amount of information needed to identify the category of a sample from $T$ after partitioning $T$ with an attribute $A$ is reduced. Through it, the ability of training sample set for attribute classification can be measured. As shown in Figure 4, the algorithm flow chart.

## 4 Integration of Computer Software Technology and Data Mining Analysis Technology in Internet Age

### 4.1 Selection of Characteristic Attributes of Computer Software

In an engineering project with large software, the defect report in software system should be described in detail and in total. Bug in software may have the following attributes: report source, author and responsible person, submission time and solution time, report content and description information, notes, priority, severity and Bug's current status. Using Bugzilla's export function, we can get the explanations of the selected attributes of all Bug report XML files as shown in Table 1, where the path relationship of the code modification function in the table is shown in Table 2.

Decision tree learning method is generally easy to transform into if-then rules, and classification rules are easy to understand. After the above analysis, we use ID3 decision tree learning method to construct classifier, which can achieve better classification effect in theory. Although the attributes of continuous values can be dealt with in decision tree learning algorithm, the learning efficiency and classification accuracy can be improved by discretising their values. The selection of discrete intervals is shown in Table 3. The values of each attribute are mapped to 0, 1, 2.

### 4.2 Classification and Analysis of Computer Software Defects

To help software maintainers to quickly understand the running status of the current system from the software version system and defect tracking system, we classify the bug repair reports of the current software system stored in Bugzilla into two categories: fixed-Bug and potential fixed-Bug.

(1) Fixed-Bug belongs to the Bug that has been repaired by the software system. It has never been reopened in the running process of the software system. These Bugs belong to the minor defects in the process of software maintenance, but maintenance personnel need to understand these defects. When similar problems occur in the

**Table 3** Discrete Processing of Characteristic Attributes

| Quantization value Feature attribute | 0 | 1 | 2 |
|---|---|---|---|
| $A_1$ | Same person | Change to another | More than two persons |
| $A_2$ | commonly | main | serious |
| $A_3$ | Only once revised | Two revisions | More than two revisions |
| $A_4$ | Containing general errors or miswritings | Including variable or function errors | Containing class or function errors |
| $A_5$ | $F_1 \to F_3$ | $F_1 \to F_3, F_1 \to_F 2$ | $F_1 \to F_2 \to F_4$ |

**Table 4** Training Set for Discrete Processing of Characteristic Attributes

| Category Feature attribute | C1(0) | C2(1) |
|---|---|---|
| Attribute Value of A1∼ A5 | 0, 0, 0, 0, 0 | 1, 2, 0, 1, 2 |
| | 0, 1, 1, 0, 1 | 1, 1, 1, 2, 2 |
| | 1, 0, 0, 1, 1 | 1, 1, 2, 2, 1 |
| | 0, 1, 2, 1, 0 | 0, 2, 1, 2, 2 |
| | 0, 2, 1, 0, 1 | 1, 0, 1, 2, 1 |
| | 1, 1, 2, 2, 0 | 1, 1, 2, 1, 2 |

running of the system, they can be repaired promptly and locate the source code of Bug.

(2) Fixed-Bug belongs to Reopen, and the modification cycle is long, and the repairman is changed many times. This kind of Bug belongs to potential Bug, which may cause problems at any time during the operation of the software system. At the same time, such bug fixes may bring new defects, requiring the Bug Report to be opened again. So software maintainers should pay more attention to potential fixed-Bug, locate each repair version through SVN and Bug number, and analyse whether new defects can still be introduced.

### 4.3 Experiments and analysis

(1) Experimental data

By using SAX parser based on event model to parse the generated XML file, we get 600 change history records after analysis. We select two kinds of small training sample sets from 600 records to train the classifier to form a pattern library, which is convenient for us to predict future defects and potential defects, as shown in Table 4. The decision tree generated by the training sample set is shown in Figure 5, where C1 and C2 in the leaf nodes represent defects and potential defects respectively.

(2) Experimental result

In this paper, we filter out 60 records from 600 bug reports and mark their categories. We use the generated decision tree to test the classification. The test results are shown in Table 5. The classification accuracy of the two kinds of defects is 85.7% and 63.6% respectively, and the overall classification accuracy is 81.7%. The accuracy of classification is shown in Table 5.

**Table 5** Classification accuracy of decision tree

| Category | Sample size | Number of correct classifications | Accuracy rate |
|---|---|---|---|
| Defect | 49 | 42 | 85.7% |
| Potential defects | 11 | 7 | 63.6% |
| Total | 60 | 49 | 81.7% |

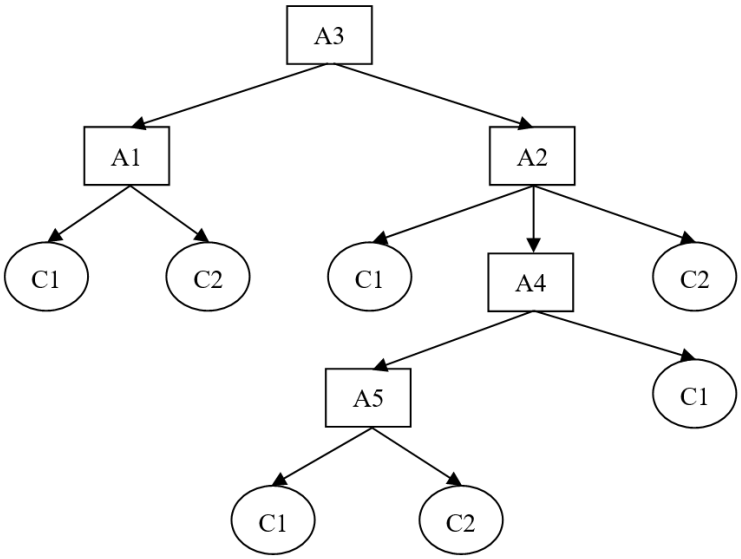



**Fig. 5** Decision Tree Generated by Defect Report

Bug reports are categorized into closed Bug reports: Bugs and potential Bugs (which have been modified several times and may be opened again). Software maintainers need to understand these Bugs of the system. For normal Bug maintainers, they only care about Bug number and Bug number annotations of source code repair in the SVN library, which can quickly locate the changed source files. For potential defective Bugs, maintainers need to focus more on these Bugs. Once they are opened again, they can locate them in the SVN Library in time. Source code file changes location.

## 5 Conclusion

With the progression of data mining technology and the expansion of data in software engineering knowledge base, researchers of health information service system have paid more and more attention to the research of algorithms suitable for software knowledge base mining. First, this paper gives a brief overview of data mining technology and computer software technology, and uses decision tree graph mining algorithm to mine functional drawings of software system definition classes, and then adds source code annotations to the relevant call relationships. In this way, software developers can quickly understand the system architecture and the associated modification of system source code files based on the function call dependency graph with source code annotations help software maintainers understand the current defect status of the system, and timely discover the change of potential introducing defects., Potential bugs in the code can be detected through these dependency graphs. This paper lays a foundation for the deep integration of computer software technology and data mining technology in the Internet era.

## References

[1] Guo M. A Study on Data Mining of Digital Display Performance of Brand Advertisement. Wireless Personal Communications. 4(2018)1-11.
[2] Khedr A E, Idrees A M, Hegazy E F, et al. A proposed configurable approach for recommendation systems via data mining techniques. Enterprise Information Systems. 12(2) (2018)196-217.
[3] Ren Q D E J, Na L. Research on Data Mining Algorithm of Meteorological Observation Based on Data Quality Control

Algorithm. Wireless Personal Communications. 102(7) (2018)1-13.
[4] Jin M, Wang H, Zhang Q. Association rules redundancy processing algorithm based on hypergraph in data mining. Cluster Computing, (4) (2018)1-10.
[5] Zhang K, Shen C, Wang H, et al. Cluster computing data mining based on massive intrusion interference constraints in hybrid networks. Cluster Computing, (4) (2018)1-9.
[6] Lasfar M, Bouden H. A method of data mining using Hidden Markov Models (HMMs) for protein secondary structure prediction. Procedia Computer Science, 127(2018)42-51.
[7] Idhammad M, Afdel K, Belouch M. Distributed Intrusion Detection System for Cloud Environments based on Data Mining techniques. Procedia Computer Science, 127(2018)35-41.
[8] Mobasheri A, Huang H, Degrossi L C, et al. Enrichment of OpenStreetMap Data Completeness with Sidewalk Geometries Using Data Mining Techniques. Sensors, 18(2) (2018)1-16.
[9] Tang W J, Yang H T, Sciubba E. Data Mining and Neural Networks Based Self-Adaptive Protection Strategies for Distribution Systems with DGs and FCLs. Energies, 11(2) (2018)426.
[10] Khosravi V, Ardejani F D, Yousefi S, et al. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. Geoderma. 318(2018)29-41.