The Implementation of a Learning Guidance Model Based on Attributes Correlation

Liyun Xiang\textsuperscript{a,b,*}, Zhiyi Fang\textsuperscript{a,*}, Hongyu Sun\textsuperscript{a}

\textsuperscript{a}College of Computer Science and Technology, Jilin University, Changchun 130012, China
\textsuperscript{b}Department of Computer Science and Technology, Changji University, Xinjiang 831100, China

Abstract

Decision tree classification algorithm is one of the effective tools to realize "intelligent" of Intelligent Guidance System. It can receive precise classification though the analyzing and mining of the data. It also has another positive characteristic that the decision tree and set of production rules are simple and efficient computing. This paper proposed the "Intelligent Guidance System" and introduced the main modules of the system. This paper also proposed the new-C4.5r algorithm based the comparing of the ID3 algorithm and C4.5 algorithm and the requirements of the individual educations. The new algorithm based on the relative of the rules attribute, simultaneously, the model of evaluation system of Calculus has been established. Experiments show that new-C4.5 algorithm is better than C4.5 algorithm on the aspects of run-time, the size of rule sets and overhead of the production rules.

Keywords: Web Data Mining; Intelligent Guidance System; Decision Tree Algorithm

1 Introduction

With the rapid development of the network, education also benefits from it, so online education becomes a hot topic today, and it is also a field which is worthy of further development. With the rapid development of educational theory the contradictions of the traditional organization of education and the educational theory are obvious. After 1980s of the 20th century, constructivist learning theory which is based on cognitive learning theory becomes popular in the teaching field and going to be the main theory of the international education reform. Constructivism emphasizes learner-centered learning environment. Further more, Strong support provided by the latest achievements of contemporary information technology makes constructivist learning theory as an important theoretical basis for online education. Internet education is the whole process composed of organization of teaching, implementation of teaching, interactions which is based on modern education thoughts and utilized by multimedia. The purpose of Internet education are designing, developing, managing and making evaluation of the learners and learning resources in order to promote the effectiveness of the learners learning. It is an amendment to the traditional

*Corresponding author.
Email addresses: liangliy09@mail.jlu.edu.cn (Liyun Xiang), fangzy@jlu.edu.cn (Zhiyi Fang).
education which is so-called "classroom teaching" so that the "individualized education" gets a better implementation of teaching principles and real implementation. Intelligent technologies, such as data warehouses and data mining techniques, are applied in the Internet education to make a better implementation of teaching principles and real implementation. Nevertheless, it also has some disadvantages, such as the delayed feedback of the student’s learning situation. Because the system which is knowledge faced can only provides Synchronous or asynchronous Materials. "Intelligent guidance system" comes into being on the premises of the demands on Interaction between Teaching and Learning and Personalized learning characteristics. It mainly simulates the traditional teaching principal of human beings, which can receive quick feedbacks in the teaching process and forms the individual learning system ultimately. The quality of both teaching and learning will be remarkably improve by providing Internet learning with intelligent guidance.

2 Web Data Mining

Great achievements have been made in the domain of Internet learning and E-learning. Great amount of teaching resources have been accumulated with the quick development of the Internet education. Simultaneously, how to use these resources effectively for individualized education becomes such an important issue that have to be dealt with immediately. The issues resolving the intelligent teaching is the core of "Intelligent Guidance System” and the destination of the system is realizing the individualized teaching and learning by taken each leaners characteristics and hobby into consideration. However, intelligent guidance and specification on each individual learner is hard to achieve because of the system can not satisfy all requirements of different learners simultaneously, such as the accumulation of the courses, the careless management to numerous different kinds of learning resources, the lack of timely feedbacks, lack of effective evaluation systems and the mechanisms of the teaching guidance. Learner model is established by use of different data mining techniques and the analysis to the information on the self-learning learners. The learning path is mined by the path-mining module. Simultaneously, C4.5 algorithm is applied in the learning-evaluation subsystem. So, the effect of autonomous learners learning has been improved, the management effect of individual teaching has been enhanced, the strong foundation of individual teaching is provided.

3 Model Design

3.1 The System Model of Intelligent Guiding System

Intelligent Guidance System was composed of the following modules: information collection module, personalized analysis and evaluation module, personalized scheduling module, learning tools module, management tools module, domain knowledge, students information, teaching rules and teaching strategy etc. [1][2][3][4]. The structure of the system are shows in Fig. 1.

Information collection is responsible for collecting the requirements from different kind of users and tracing the action of each user by taken the unit of domain knowledge point as the basis. At the same time, it also takes in charge of storing the information researched into database. Both the learning information and web catalog information were collected and used as the input data
source to personalized analysis and data mining process. With the help of those two kinds of advanced technologies, both the common characteristics that are shared between multiple users and individual characteristics of single learner were abstract and stored into the rule sets database. Teaching strategy mainly concentrated on recording the teaching strategies which are summarized by the teaching experts according to their teaching experiences. The approach to record the strategy is done by utilizing the knowledge representation with different weights assigned. These rules will transfer to the personalized schedule module when learners logged into the system with authentication. Then the system will schedule the tools and resources according to the rules and provide the personalized learning environments for each distinctive learners. Through above procedure, the requirements of the individual learning are met. Evaluation system will modify the corresponding parameters and the credibility of the system rules according to the learning affects of each learner. The rules after several times of proliferation and modification will better reflect the actual requirement of each distinctive user. At the same time, remarkable affect will be achieved with more accurate credibility of rules.

![Fig. 1: The model of intelligent guiding system based on the network](image)

### 3.2 Domain Knowledge Base

Domain knowledge comprised the teaching concerned knowledge and the teaching resources available. It is mainly composed by related teaching domain knowledge and the relationships among those knowledge points. It is represented by the combination of relationships between one knowledge point and another one. Domain knowledge was composed of two sun-modules: knowledge points attributes library and knowledge tree library. Knowledge points library, as the basic unit, is the set of the teaching contents. Such as knowledge points may be facts, concepts, rules, theorems, definitions and problem solving methods. Each knowledge point was directed to a number of relevant teaching materials, Such as multimedia material, question database, test paper mate-
rial, coursewares, case studies, literature and so on. Both the characteristics and the relationship among those points were described by attributes in this system. Such as the understanding of the knowledge attribute is used to mark points in the course of importance, the knowledge of the relationship between the points reflect the knowledge of property between points relationship. Attribute related knowledge points forms an unique network structure, Bigger degree of knowledge points will be gathered in a subset, and the knowledge points will be formed to courses, units, chapters, knowledge point, materials. As shown in figure 2, according to the figure mentioned the virtual layers was represented by dotted line who is composed of courses, chapters, and these three layers was represented by the knowledge point code as shown in table 1. The relations among those knowledge points were organized by a tree data structure and stored in the knowledge library.

![The model of domain knowledge hierarchy](image)

**Fig. 2:** The model of domain knowledge hierarchy

<table>
<thead>
<tr>
<th>course code</th>
<th>chapter number</th>
<th>section number</th>
<th>knowledge point number</th>
</tr>
</thead>
</table>

### 3.3 The Design of Higher Mathematic Discipline Personalized Evaluation Module

In the "Intelligent Guidance System", the key step is establishing the teaching rules according to the students’ information which was left automatically in the information library after the learners visit the knowledge library. Those records that are time continuous in nature were converted into time sequences that are time partitioned in nature by means of data purification and transformation. The behavior of learners will be mined by using sequential ming methods. This will further enrich the teaching strategies database. This will provide each users with personalized learning guidance and makes the concept of ”personalized education” come into practice. For example after a period of data mining, the system can find the best sequence of the certain knowledge point compared with the grade of the learning. And these best sequences can be used as the default learning path or recommending learning path. As shown in Fig. 3, decision tree classification algorithm was used in the realization of personalized analysis and evaluation modules. The teaching strategy was optimized by analyzing and evaluating the learning results.
4 Algorithm Design

4.1 C4.5 Algorithm

Decision tree is commonly used as data mining algorithm tools. Decision tree classification algorithm uses top-down, divide and conquer recursively to divide the data into several subsets and build the corresponding classification model in a tree-based topology [5]. The application of Decision Tree Algorithm in Intelligent Guidance System is the key to reflect its intelligence. Quinlan proposed constructing a decision tree ID3 algorithm in 1979. In ID3, the choice of decision node attributes is determined by the concept of entropy in information theory. Through the property of information information to gain the maximum (or the maximum entropy compression) to create the decision tree and the property of this node can guarantee the decision tree has the least branch number and smallest redundancy [6]. Later, he proposed C4.5 which is more advanced. C4.5 algorithm amends the pruning algorithm ID3, and is good for high branching property, value-type properties and properties with null values. In this paper, based on the c4.5 algorithm, the new-C4.5r decision tree algorithm is proposed. New-C4.5r algorithm is based on the correlation of simplification rules, and uses this algorithm to build the model of Higher Mathematics Personalized Learning Evaluation System.

4.2 Improvement of Algorithm

When applying C4.5 algorithm to classify some unknown samples, the system may encounter the "over-fitting" problem. As a result, it is necessary to simplify the samples before they are classified. The following are the processes of the improved algorithm, named new-C4.5r:

- Use C4.5 tree to construct a complete decision tree $T$.

$T$ will be converted to the rule set $R$. The rule $r$ corresponds with a path from the root node to a leaf node in the $T$.

$R:r_i — if \text{Cond}_1 \land \text{Cond}_2 \land \ldots \land \text{Cond}_n \text{ then class } C_x.$

Simplify each rule $r_i$ of $R$ as following:

$i = 1;

\text{While } (i \leq n)$
\[
\begin{align*}
    t_{i,j} &= P(\text{Cond}_{i+1} \land \text{Cond}_i); \\
    \text{if } (t_{i,i+1} \geq \lambda) \text{ then} & \quad i = i + 1; \\
    \text{else} & \quad \\
    \{ & \quad \text{delete } \text{Cond}_{i+1} \land \ldots \land \text{Cond}_n \text{ in } r; \\
    & \quad \text{break;} \\
\}
\end{align*}
\]

In the above process, introduce the parameter \( \lambda \) as the threshold value of \( P(\text{Cond}_{i+1} \land \text{Cond}_i) \). It is similar to the concept of the Minimum Support in the Association Rules. Its default value is 0.15\%\[7\]. The value of \( \lambda \) is controlled to eliminate the over-fitting part of the rules. Merge and simplify the same rules in \( R \), and get a new rule set \( R' \).

Establish an attribute-associated matrix \((T_{vs})_{n \times n}\).

- If \( t_{vs} = 0 \), Attribute \( A \) and Attribute \( B \) are irrelevant.

- If \( t_{vs} = 1 \), Attribute \( A \) and Attribute \( B \) are relevant.

Simplify each rules \( r' \) of \( R' \) as following:

\[
i = 1; \\
\text{While } (i \leq n) \\
\{ \\
    \text{for } (j = j + 1; j < n; j + +) \\
    \{ \\
    \quad \text{if } (t_{vs} = 0) \text{ then continue;} \\
    \quad \text{if } (t_{vs} = 1) \text{ then} \\
    \quad \{ \\
    \quad \quad \text{Calculate the condition probability in the training set: } P(\text{Cond}_j|\text{Cond}_i), P(\text{Cond}_i|\text{Cond}_j); \\
    \quad \quad \text{if } (P(\text{Cond}_j|\text{Cond}_i) \geq P(\text{Cond}_i|\text{Cond}_j)) \text{ then Tag and eliminate } \text{Cond}_j; \\
    \quad \quad \text{if } (P(\text{Cond}_i|\text{Cond}_j) > P(\text{Cond}_j|\text{Cond}_i)) \text{ then Tag and eliminate } \text{Cond}_i; \\
    \quad \quad \text{break;} \\
    \quad \} \\
    \} \\
    i = i + 1; \\
    \text{While } (\text{Cond}_j \text{ is tagged}) \\
    \quad i = i + 1; \\
\}
\]

In this process, at first, the correlation between attribute \( A_v \) and \( A_s \), which are belong to \( \text{Cond}_i \) and \( \text{Cond}_j \) respectively, is judged. If \( A_v \) relates to \( A_s \), \( \text{Cond}_i \) and \( \text{Cond}_j \) in rules will be kept, otherwise, both conditional probabilities are calculated, and \( \text{Cond}_i \) or \( \text{Cond}_j \) will be eliminated according to the confidence. The same rules of \( R' \) are merged and simplified, and a new rules set \( R'' \) is obtained \[8\].
5 Experimental Result Analysis

In this paper, against a set of data, tests were made separately to compared the C4.5 algorithm and new-C4.5 algorithm. The test parameters comprised by running time, classified right ,the number of rules , the size if the rule set. The results are shown by table 1. The rules set of the new-C4.5 algorithm is abstracted from the set of C4.5 algorithm by the means of simplifying the similar rules. Experiments data certified that new-C4.5r algorithm has obvious advantage in running time on the premises of the similar accurate of the classification.

Table 2: Comparison of C4.5 rules algorithm and new-C4.5r algorithm

<table>
<thead>
<tr>
<th></th>
<th>the size of training set</th>
<th>running time (sec)</th>
<th>classified right (%)</th>
<th>the number of rules</th>
<th>the size of the rule set</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5r algorithm</td>
<td>30000</td>
<td>82</td>
<td>88.6</td>
<td>46</td>
<td>198</td>
</tr>
<tr>
<td>new-C4.5r algorithm</td>
<td>30000</td>
<td>49</td>
<td>86.4</td>
<td>29</td>
<td>124</td>
</tr>
</tbody>
</table>

6 Conclusion

The construction of the "Intelligent Guidance System" has magnificent meanings of Internet education the technology of data mining and its application play a crucial role in the aspect of "Intelligent". This paper proposed a simple-production based new-C4.5r algorithm though analysis and research of the C4.5 algorithm. Experiments show that new-C4.5 algorithm is better than C4.5 algorithm on the aspects of run-time and the size and overhead of the production rules however, it must be improved in future ,such as the execute time, the scale of the production rule, an objective evaluation, in order to apply to the "Intelligent Guidance System" efficiently.

References