Study of the Learning Model based on Improved ID3 Algorithm

Ding Rongtao
470,Binwen Road
Hangzhou,China
86-130-18961726
rongtaoding@gmail.com

Ji Xinhao
470,Binwen Road
Hangzhou,China
86-133-05715009
jxh@zjvcc.cn

Zhu Linting
470,Binwen Road
Hangzhou,China
86-135-88818168
annetty@163.com

Ren Wei
470,Binwen Road
Hangzhou,China
86-133-25815303
hzshenyi@vip.sina.com

Abstract

The network learning behavior intelligence analysis system can collect the information of learner's psychology, behavior, methods and effectiveness in the learning process, and classify learners by using the ID3 algorithm based on the internal factors and personality characteristics of learners that influence the learning effect. In order to correct the shortcomings that the ID3 algorithm more inclined to the attributes that have more values in the classification process, we introduce user interest, which used to distinguish the dependence between different information attributes. At the same time, we introduce parameters to reduce the redundancy between attributes, and accelerate the pace of information entropy reducing, then construct a general, expandable senior vocational student model in the intelligence-learning environment.

Keywords

Decision tree; ID3 algorithm; Learner model

1.INTRODUCTION

The education expert system that has realized personalized teaching becomes the key when focused on the development of network education. It is student-centered, and it develops learning strategies, distributes teaching resources and constructs the individual virtual learning environment for learners on the basis of certain rules and the character of learners. Individualized education involves many theories such as modern pedagogy and psychology; the core issue is how to construct student-learning model based on the learner's cognitive level, character, motivation and fancy. In this paper, we select the scientific classification system of students learning ability and learning behavior by using the decision tree algorithm in the data mining, and construct student-learning model.

Conference'04, Month 1–2, 2004, City, State, Country.

2.Related background

2.1 Learning behavior analysis system

Learning model is the classification model according to the subjective will and the objective ability in the process of students learning [1]. Constructing learning model is not a simple classification, but must classify and synthetically

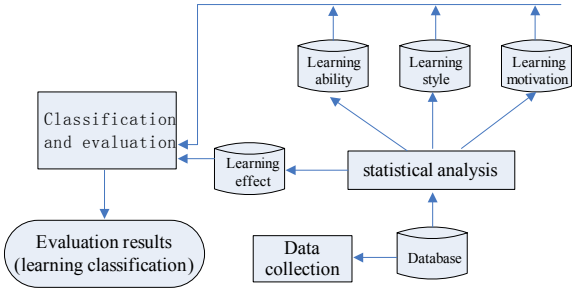

Figure. 1 Network learning behavior analysis system

evaluate learner's learning ability, learning mode and motivation firstly. In order for the correct analysis and evaluation of learner's learning model, we have devised a network learning behavior intelligent analysis system to collect data and mine data, then realize the classification and evaluation finally, as shown in figure 1. The system composed of data collection module, statistical analysis module, and classification and evaluation module. The data collection module is mostly used for the collection and quantification of learning ability, behavior, strategies and tendencies, which impact learner's learning. Statistical analysis module embedded process the collected data by data mining technology, and explores the personality characteristics. The results of statistical analysis can be the basis of classification and evaluation module, and provide the foundation for the establishment of the "learner characteristics - results" model. Classification and evaluation module is further mining classification of the history data of network learning behavior, and analyze and evaluate the relationship between personal learning behavior and learning effect.

2.2 ID3 algorithm

ID3 algorithm is a typical decision tree algorithm. It analyzes known types of objects according to a fixed set of attribute or characteristic, and produces a decision tree, and then the decision tree put all the objects in the correct classification [2]. It uses the concept of mutual information when choosing important characteristics, forms the decision tree by using the subset of training examples, and excerpts mutual information as the discriminance. First, find out factors that have best sense, and divide the data into several subsets, and then each subset can be divided by the factors that have best sense, till all subset contains the same type of data, thereby result in a decision tree.

ID3 algorithm has clear theory, simple technique and strong learning ability; it is suitable for processing mass resources distribution issues. But ID3 algorithm has its drawbacks: ① the calculation of mutual information depends on the characteristics that have much eigenvalue. ② there is a hypothesis if we set the mutual information as a feature selection method. That is the proportion between positive examples and negative examples in training example subsets should be the same with the proportion in the real problems. But it cannot be guaranteed the same under normal circumstances, and there is deviation when calculate the mutual information in training set. ③ ID3 algorithm is sensitive to the noise (the errors in training sets). ④ ID3 decision tree will be changed along with the increasing of training sets. And it is inconvenient to the growing of training examples.

3. Improved ID3 algorithm

Traditional ID3 algorithm chooses attributes, and often tend to choosing the attributes that get more values, because the weighted sum method makes the classification of examples set tend to the metadata group that discarding small data group, but the attribute has more properties is not always optimal one. The attributes in the learning model building process include the knowledge level of originally subject in learning ability database, the multiple factors of learning mode in learning mode database, and the learning motivation classification in learning motivation database. The final decision tree classification results are not certainly consistent with the actual situation according to the traditional ID3 classification because there are many types of attributes.

3.1 Introduce the user interest α

In the decision tree established by increasing user interest, the information entropy corresponded to the root node is the largest. Along with the construction of the decision tree, information entropy gradually decreased until the entropy of leaf nodes turn to zero (i.e. all objects of a node in the same category). Therefore, it is hoped that each choice of testing attribute can reduce the entropy at the presto speed, and then make every branch of the decision tree as short as possible and eventually build a smaller tree. It is the purpose of ID3. The traditional ID3 algorithm does not take into account the influence of the relationship between attributes on the attributes choosing, and results in the choice of redundant attributes that have little meaning or no significance to the real classification. Algorithm demands the maximal relationship between the selected attribute and the genus (i.e. the information gain in ID3 algorithm), and the minimal relationship with the used attributes in the same branch (interactive information) [3]. This will avoid the choice of redundancy attributes, and accelerate the pace of entropy reducing and thus build a better tree.

In order to distinguish attributes importance, we introduce the user interest when calculating the information entropy to distinguish the dependence of attributes. The user interest α ($0 \le \alpha \le 1$) to be known as the user interest to uncertain knowledge, and it is determined by the decision-makers according to the prior knowledge or area knowledge. It is a vague concept, usually referred to certain prior knowledge, including area knowledge and expert advice. And in the study of decision tree, it is referred to the factors that influence the generation and selection of the decision tree rules except the examples set used for the formation and modification of the decision tree in its training process.

Suppose that a training examples set is X, the purpose is to divide the training examples into n classes, recorded as C = (X1, X2, … Xn). On the assumption that the number of ith training examples is | Xi | = Ci, the probability that an example belongs to this training examples is P (Xi). If we choose the attribute A to test, with a set of properties $a_1, a_2, a_3, \ldots a_i$, the number of examples that belonged to the ith category when $A = a_j$ is Ci_j.

$$P(Xi : A = aj) = \frac{Cij}{|X|} \tag{1}$$

The value of $P(Xi : A = aj)$ is the probability that the test attribute A belongs to the ith category. Y_j is the examples set when $A = a_j$, then the degree of uncertainty to the decision tree classification is the entropy of the training examples set to attributes A:

$$H(Y_j) = -\sum P(X_i \mid A = a_j) \log_2 P(X_i \mid A = a_j) \tag{2}$$

We increase the user interest α when calculating the taxonomic information entropy of each leaf node X_j when $A = a_j$ extended from attribute A, and then strengthen the label of important attribute, and reduce the label of non-important attribute. The formula as follows:

$$H(X \mid A) = \sum [P(A = a_j) + \alpha] H(X_j) \tag{3}$$

The information provided by attributes A for classification (the information gain of attribute A) is:

$$I(X : A) = H(X) - H(X \mid A) \tag{4}$$

3.2 Introduce the parameters λ

The formula four will be extended into:

$$\Delta = I(X; A_i) - \lambda \sum_{Ai \in S} I(A_i; A_s) \tag{5}$$

S is the attribute already used in ancestor node, $\sum_{A_i \in S} I(A_i; A_s)$ is the sum of relationship between the attributes that to be selected and the attribute already used in ancestor node, $\lambda(\in [0,1])$.

3.3 Improved calculation steps

（1）Initialization：

$F \leftarrow$"all attribute″,

$S = \varnothing$,

$X \leftarrow$"all training examples″,

$I[n,n] \leftarrow \underset{1 \le i,j \le n}{I}(A_i; A_j)$;

（2）Creating the root node：

① To all of the attributes, we calculate the interactive information $I(X; A_i)$ between each attribute and the genus, and select the attribute A_b that has the best relationship as the root node;

② $F \leftarrow F - \{A_b\}$, $S \leftarrow \{A_b\}$;

（3）We can set the node as the leaf node if all the objects in X belong to the same type, and mark it to the corresponding categories and return. If all the attributes have been set, we can calculate the believe level CF, and select the maximal value as the category label of this leaf node and return;

（4）We set the node to be the test node if it does not meet the conditions upward; its production process is as follows:

① To each attribute A_i that has not been used in the same branch, calculate the value of $\Delta = I(X; A_i) - \lambda \sum_{A_i \in S} I(A_i; A_s)$, and select the attribute A that has the maximal value as the test attribute;

② Divide X into $X_1, X_2, X_3, \ldots\ldots X_N$ according to the value of A, and produce new nodes $t_1, t_2, t_3, \ldots\ldots, t_N$, then mark the attribute value to each edge;

③ $F \leftarrow F - \{A\}$, $S \leftarrow \{A\}$;

④ Replace X with $X_1, X_2, X_3, \ldots\ldots X_N$, and repeat step (3).

4.Learning Model Implementation

4.1 Data collection and pretreatment

The major sources of data collection in the following ways:

(1) Testing evaluation: include the test of learner's original subject knowledge level before the courses, the staggered self-test in the curriculum learning process, and the comprehensive test of learning effect in the end of the course.

(2) The questionnaire: in order to understand the learners learning strategy, learning behavior and motivation. We have devised a questionnaire to investigate the vocational students.

(3) Web log visitation: we can understand learners learning tendencies by visiting the web log, and judge the master degree of the learning tools in network learning activities.

The result of data collection is a large number of discrete information, we adopt statistical analysis module to quantify these discrete data, and then store them in the learning ability database, learning styles database and learning motivation database.

The result of test before the courses is discrete data; we can quantify it into three grades.

$$\chi = \begin{cases} 1, & 85 \le \alpha \le 100 \\ 0.5, & 60 \le \alpha \le 84 \\ 0, & \alpha \le 59 \end{cases}$$

χ is the original academic knowledge level , α is the test results.

We classify the questionnaire subjects according to the test objectives, and obtain the personality, cooperation, interaction, practicality, inquiry, originality, sense, adaptability and dependence in learner's motivation, knowledge ability, and learning styles, the value in the range of [0. 0-1.0]. Statistics results are shown in Table 1.

To facilitate the calculation of classification, we quantified the

Table 1.Statistics results

Student number	1	2	…
Personality	0.2	0.4	…
Cooperation	0.6	0.5	…
Interaction	0.7	0.5	…
Practicality	0.5	0.4	…
Inquiry	0.5	0.3	…
Originality	0.8	0.2	…
Sense	0.4	0.5	…
Adaptability	0.6	0.4	…
Dependence	0.6	0.5	…
Learning motivation	0.2	0.5	…
Knowledge level	0.6	0.5	…

data further by the range:

$$x = \begin{cases} low & 0 \le \alpha \le 0.3 \\ normal & 0.4 \le \alpha \le 0.7 \\ high & 0.8 \le \alpha \le 1.0 \end{cases}$$

Quantized data is shown in Table 2.

We set higher weight to the web log data when the statistical results of the web log differ from the survey results, and then calculate the weighted average with the data from report.

4.2Building the Decision Trees

For obtaining the information gain of each attribute, we can calculate the expectation information needed by certain samples classification by using the formula (4).

For example, we get the user interest α = 0.3, and build the decision tree according to the learning style information. Choose the value of personality attributes, and calculate the attribute information entropy.

$$H(\frac{X}{Personality}) = (\frac{250}{382}+0.3)*(-\frac{180}{250}\log_2\frac{180}{250}-\frac{70}{250}\log_2\frac{70}{250})+0+(\frac{250}{382}+0.3)*(-\frac{100}{250}\log_2\frac{100}{250}-\frac{150}{250}\log_2\frac{150}{250}) = 0.973$$

Similarly, we have tested user interest of other relevant

Table 2. Quantized data results

Student number	1	2	…
Personality	low	normal	…
Cooperation	normal	normal	…
Interaction	normal	normal	…
Practicality	normal	normal	…
Inquiry	normal	low	…
Originality	high	low	…
Sense	normal	normal	…
Adaptability	normal	normal	…
Dependence	normal	normal	…
Learning motivation	low	normal	…
Knowledge level	normal	normal	…

attributes, and calculate the information entropy of relevant attributes, including the personality, cooperation, interaction, practicality, inquiry, originality, sense, adaptability and dependence.

$$H(\frac{X}{Interaction}) = 0.962 ; H(\frac{X}{Practicality}) = 0.933 ;$$

$$H(\frac{X}{Dependence}) = 0.925 ; H(\frac{X}{Cooperation}) = 0.942 ; H(\frac{X}{Inquiry}) = 0.946 ;$$

$$H(\frac{X}{Originality}) = 0.901 ; H(\frac{X}{Sense}) = 0.912 ; H(\frac{X}{Adaptability}) = 0.958 。$$

It is known from the comparison of the information entropy between attributes that:

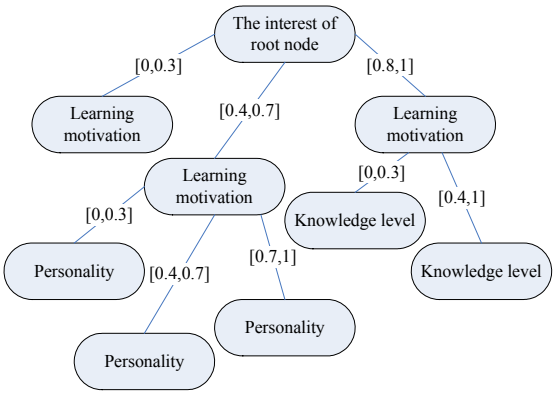

Figure2. Students decision tree classification model

$$H(\frac{X}{Personality}) \geq H(\frac{X}{Interaction}) \geq H(\frac{X}{Adaptability}) \geq H(\frac{X}{Inquiry}) \geq H(\frac{X}{Cooperation})$$
$$\geq H(\frac{X}{Practicality}) \geq H(\frac{X}{Dependence}) \geq H(\frac{X}{Sense}) \geq H(\frac{X}{Originality})$$

We classify the students according to attributes, and obtain the final decision tree by repeating this process, shown in Figure 2.

5. Learner classification result

The students can be roughly divided into four categories according to the statistics from survey and web log.

The first category:

These students lack the spirit of adventure, prefer the step by step process of learning to the challenge learning. They follow the goal demanded by teachers, and hope that every part has complete framework. Students in this category can realize self-monitoring learning, and search for some additional information resources to help themselves learning more new knowledge of the subject. In their opinion, learning is a very important way to achieve personal goals.

The second category:

They are suspicious of the goals set by others, and doubt about the successfully achievement to these objectives. They wonder whether these learning activities are helpful to the realization of their personal goals and achievement, and whether their personal values will be reflected.

The third category:

They love learning, and have self-management and self-rule abilities. They dare to take the risk and actively participate in the learning process, and search for some additional information resources on their own initiative to help themselves to learn more about the new curriculum knowledge. They are happy to discuss with other students, and are not entirely confined by the short-term tasks, objectives, schedules and deadlines, and their aims will not be influenced by foreign power, such as general operating norms, social and educational expectations.

The fourth category:

Such learners focus on the short-term goals, they are driven by the learning task and the external power. For example, they pay attention on the score, the encouragement, and general effect standards. They are unwilling to take the risk, to meet the challenge, and to pursue some goals seems difficult to achieve. They often depend on the mentors and the external resources, and also need strong reasons to support their learning.

The first category is known as compliant learners, the second category is known as treasonous learners, the third category is known as versatility learners, and the fourth category is known as hard working learners.

Most of us are hard working learners or versatility learners, who have self-learning ability, strategies and initiative, but cannot be separated from teachers.

6. Conclusion

In the learning process, the content and media display of teaching should synthetically consider knowledge, learning styles and motivation of individual learners. We can formulate right learning elements, provide appropriate learning content, identify individual differences in learning, and improve interest, value and efficiency according to learner's character. This article has designed a network learning behavior intelligent analysis system, which focused on the data collection and processing and the conformation of learning classification model. It improved the ID3 algorithm in decision tree classification, introduced α as user interest to distinct the importance between properties, and parameter λ to reduce redundancy between properties, set up a classification model to the learner's ability and learning style, then provide a reference guide for teaching students in the higher vocational schools, and also can be the evidence for the management and design in distance education.

7.References

[1]. Guoxiang Zhong, etc, A Study of Building a Student Agent Model in Intellectual Learning Environment by Using Bayesian Network [J]，Computer Science，2006 (33)：203~206

[2]. Jingfeng Guo, etc, Research on the Parallelity of Decision Tree Algorithm, Computer Engineering.2002, 28(8): 77-78

[3]. Haixun Wang,Yu P S, SSDT: A Scalable Subspace-splitting Classifier for Biased Data. ICDM 2001 Proceedings, IEEE International Conference Proceedings, 29 Nov.-2 Dec:542-549