Eng Int Syst (2020) 4: 223–227

Engineering
Intelligent Systems

Improved ID3 Algorithm Based on Intelligent Computer Distance Education

Yunxia Wu[1]*, Hengjie Zhang[2] and Xin Li[3]

[1] *School of Continuing Education, Shijiazhuang University of Applied Technology, Shijiazhuang 050081, China*
[2] *Department of Information Engineering, Shijiazhuang University of Applied Technology, Shijiazhuang 050081, China*
[3] *Department of Electrical and Electronic Engineering, Shijiazhuang University of Applied Technology, Shijiazhuang 050081, China*

At present, the distance education application system lacks intelligence, as well as innovation, and cannot provide a personalized teaching method for learners on different levels. After introducing the Iterative Dichotomiser 3 (ID3) algorithm online learners can be classified intelligently according to their inherent characteristics, to achieve targeted teaching for learners of different levels. However, the traditional decision tree ID3 algorithm has the problem of multi-value tendency, and the selection of split attributes does not work with objective facts automatically. A modified factor attribute selection method based on gray association analysis, focusing on an intelligentized target, is used to improve the properties with more values but a lower gray association degree. The sine value of a gray association degree is used as the correction factor to overcome the deficiency of the traditional ID3 algorithm when calculating the information gain of the properties. By introducing the improved ID3 algorithm into the distance education system, learners can be better classified to achieve intelligent learning guidance.

Keywords: ID3 algorithm; decision tree; gray correlation degree; correction factor; distance education system; intelligentized design

1. INTRODUCTION

Traditional distance education systems are usually centered on the system itself, and can provide only identical learning materials and tasks to different students, thereby lacking system intelligence [1]. With the continuous development of network information technology, various electronic information resources have become increasingly rich, enabling learners can to find all kinds of relevant learning resources from the Web. However, how to quickly locate users from the massive volume of information, find the information resources they need to become the key to information services has become an issue.

Modern distance education is a new, web-based mode of delivering education. However, it does not take into account the learner's capacity to study in a network environment, his/her preferred media, learning skills and/or disabilities. The current distance education websites are not designed to meet the individual learning needs of students as shown in Fig. 1. Hence, there is a need to develop distance education websites that are learner-centered rather than teacher-centered, so that individual learner characteristics are considered and teaching strategies and course content can be personalized accordingly [2].

Based on Web technology and text data mining, Saura et al. proposes the application of a content recommendation algorithm to a personalized distance education service. They use a simulation experiment to verify the effectiveness of the integrated algorithm, and then apply the tested algorithm to the construction of a website higher vocational course teaching [3]. The collaborative filtering algorithm [4] is the most widely used personalized recommendation algorithm, but the traditional recommendation algorithm considers the interests of users at different times and its recommendation results are not timely. To solve this problem, an improved time-based

*Email of Corresponding Author: wuyx1975@126.com

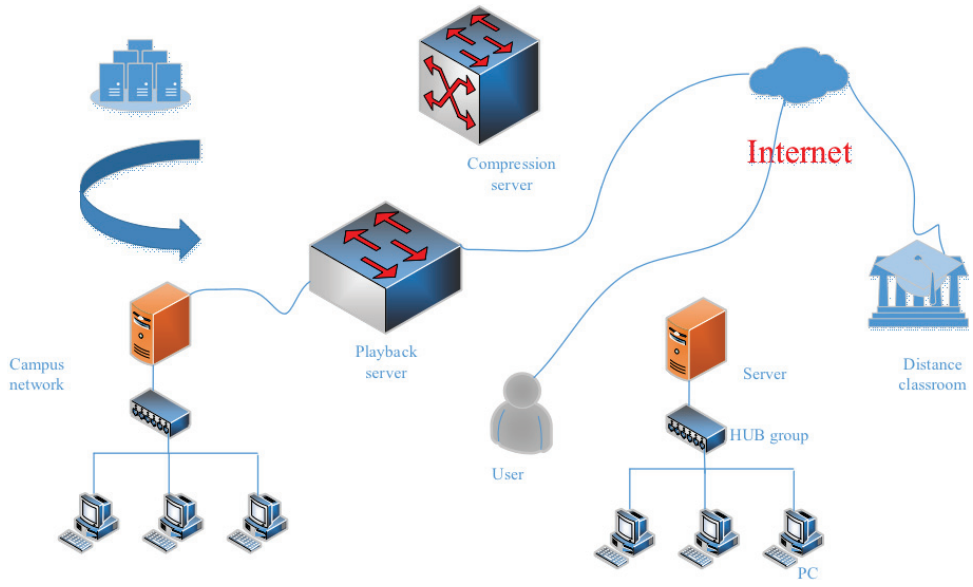

Figure 1 Framework of the distance education system.

collaborative filtering algorithm is proposed in this paper, so that users' interests closer to the collection time have more weight in the recommendation process, thereby improving the accuracy of the recommendation. To remedy the low level of personalized teaching in distance education, Azzi et al. proposes a clustering algorithm based on a rough set, and the application of the rough set reduction method to solve the learners' characteristics in the data attribute redundancy, improving the efficiency of the clustering algorithm [5], to improve the level of personalized teaching on a distance education website.

However, data mining technology can find some potential and valuable rules from the massive data of the remote teaching system, which undoubtedly provides strong support for intelligent and personalized online learning [6]. Given the disadvantages of the traditional online learning system, this paper uses the decision tree Iterative Dichotomiser 3 (ID3) algorithm to classify learners and determine their learning needs according to their test scores thereby achieving intelligent learning guidance. However, due to some defects in the ID3 algorithm, it was improved to obtain the Genetic-Based Iterative Dichotomiser (GBID) algorithm. After this improvement, both the efficiency and classification accuracy of the algorithm were greatly enhanced, so as to better realize the intelligence of a remote teaching system.

2. ID3 ALGORITHM ANALYSIS

The ID3 algorithm is a decision tree learning algorithm based on information entropy proposed by Quinlan in 1986. Information entropy is a quantitative measure of information, where information gain is the difference between two pieces of information entropies, representing the amount of information obtained after eliminating uncertainty [7]. The ID3 algorithm is a greedy algorithm; that adopts the recursive top-down, divide and conquer method to construct the decision tree. The core of this algorithm is defined as follows: when selecting attributes on nodes at all levels of the decision tree, the optimal split attribute is selected by calculating the information gain [8]. Firstly, the attribute with the maximum information gain value is selected as the root node, and then the branch is created according to the different values of the root node, which also corresponds to a divided subset. The ID3 algorithm is then recursively called on each subset to establish the node branch of the decision tree until the whole decision tree is generated.

The specific description of the ID3 algorithm is as follows [9–10]. Supposing that S is a data sample set, which contains S data. The class label attribute has n different values, where n is defined as different kinds of $C_i (i = 1, \ldots, n)$. Let S_i be the number of samples in class C_i. The expected information required by the classification of a given sample is given as

$$I(s_1, s_2, \ldots, s_n) = \sum_{i=1}^{n} p_i \cdot \log_2(p_i) \tag{1}$$

where p_i is the probability that any sample belongs to C_i and is estimated by $\frac{S_i}{S}$.

Let attribute A have k different values $\{a_1, a_2, \ldots, a_k\}$, applying attribute A divides S into k subsets $\{S_1, S_2, \ldots, S_k\}$, where S_j contains these samples in the set S that has the value a_j in attribute A. If A is chosen as the test attribute (the best split attribute), then these subsets correspond to the branch that grows from the node containing the set S. S_{ij} is the total number of samples of class C_i in the subset S_j. According to the entropy, or expected information divided into subsets by A, the formula is given [11].

$$E(A) = \sum_{j=1}^{k} \frac{s_{1j} + s_{2j} + \ldots + s_{nj}}{s} I\left(s_{1j}, s_{2j}, \ldots, s_{nj}\right) \tag{2}$$

The term $\frac{s_{1j}+s_{2j}+\ldots+s_{nj}}{s}$ acts as the weight of the j-th subset, and is equal to the value of the subset, namely, the property A belonging to the number of samples in a_j is divided by the total number of samples in S. The smaller the entropy, the higher the purity of the subset. For a given subset of S_j,

$$I\left(s_{1j}, s_{2j}, \ldots, s_{nj}\right) = -\sum_{i=1}^{n} p_{ij} \cdot \log_2\left(p_{ij}\right) \tag{3}$$

where $p_{ij} = \frac{s_{ij}}{|s_j|}$ is the probability of sample in S_j belonging to the category C_i. The achieved information gain in an attribute A branch is

$$G(A) = I(s_1, s_2, \ldots, s_n) - E(A) \tag{4}$$

$G(A)$ is the expected compression of entropy due to knowing the value of the attribute A. The information gain of each attribute is calculated by this algorithm. The attribute with the highest information gain is selected as the split attribute of the given set S. A node is created and marked with that attribute, a branch for each value of the attribute is created, and the sample is divided according to reference [12].

As a classical construction algorithm of the decision tree, the ID3 algorithm has the following advantages: the search space is a complete hypothesis space, the objective function must be in the search space, and there is no danger of having no solution. The basic theory of the algorithm is relatively clear, and the concept of information gain is used in attribute selection. Each branch of the decision tree corresponds to a classification rule, which can generate easily-understandable IF-THEN classification rules. Therefore, the resulting classification rules are intuitive and easy to understand. However, in recent years, scholars have also found shortcomings in the research on the ID3 algorithm. When calculating information gain, they tend to choose an attribute with multiple values, which is not always reasonable, as the attribute with more values is not the optimal attribute in many cases. This means that during tree construction, it is necessary to sort and scan the data set from top to bottom many times, so the processing efficiency of the algorithm is low [13].

3. CORRECTION FACTOR BASED ON GRAY CORRELATION ANALYSIS

Gray correlation analysis refers to the method of quantitative description and comparison of the development and change trend of a system. It is conducted to determine whether the reference data column and several comparison data columns are closely related by examining the similarity of their geometric shapes, which indicates the degree of correlation between curves [14]. Firstly, the correlation coefficient of the two is obtained; the degree of correlation is obtained from the correlation coefficient, the correlation degree is sorted and analyzed according to the degree of correlation, and the conclusion is then drawn. Through some methods, the gray correlation analysis can find the changing trend of two factors in the system to judge the degree of correlation between them. If two show a great amount of change, they can be considered to be more correlated; otherwise, the correlation between the two is small. Therefore, gray correlation analysis provides a quantitative measurement for the development and change trend of a system, which is very suitable for the analysis of dynamic processes [15]. Gray correlation analysis involves calculating the degree of correlation between feature attributes and classification attributes. The sine value is taken as the correction factor to recalculate the Gain Value of attributes in the ID3 algorithm, to solve the multi-value bias problem of the ID3 algorithm. The method used to calculate the correction factor using gray correlation analysis is as follows: supposing that the number of samples in the training data set T is n, the category attribute is denoted as C, and the characteristic attribute is denoted as m, respectively denoted as $X_i(i = 1, 2, 3, \ldots, m)$, then according to the gray system theory, compare the relationship between the attributes and calculate the correlation degree of the two. Therefore, it is assumed that the category attribute values of n samples constitute a gray sequence: $C = \{C(1), C(2), \ldots, C(n)\}$; The characteristic attribute values of n samples also constitute a gray sequence: $X_i = \{X_i(1), X_i(2), \ldots, X_i(n)\}(i = 1, 2, \ldots, m)$. The gray correlation coefficient of characteristic attribute sequence X_i and category attribute sequence C at the k-th point (sample) is defined as

$$\zeta_i(k) = \frac{\min\limits_i \min\limits_k |C(k) - X_i(k)| + \zeta \max\limits_i \max\limits_k |C(k) - X_i(k)|}{|C(k) - X_i(k)| + \zeta \max\limits_i \max\limits_k |C(k) - X_i(k)|} \tag{5}$$

where the term $\min\limits_i \min\limits_k |C(k) - X_i(k)|$ is the two-layer formula, and the minimum value of the absolute difference is calculated. The first layer is the minimum value of the absolute difference of

$$CF(A) = \begin{cases} \sin[r(A)], & \text{if attribute } A \text{ has multivalue tendency} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

each point on the attribute sequence $C(k)$ and the characteristic attribute sequence $X_i(k)$, respectively, and then the minimum value is selected from these minimum values. The term $\max\limits_i \max\limits_k |C(k) - X_i(k)|$ is calculated by taking the maximum value of the absolute difference of the two-layer formula. The first layer is calculated by taking the maximum value of the absolute difference of each point on the attribute sequence $C(k)$ and the characteristic attribute sequence $X_i(k)$, respectively. The term $|C(k) - X_i(k)|$ is the absolute difference between each point on the attribute sequence $C(k)$ and each point on the attribute sequence $X_i(k)$. ζ is the resolution coefficient between 0 and 1, usually 0.5. Finally, by combining the correlation coefficients of all feature attribute sequence points (samples), the gray correlation degree of the whole attribute sequence $C(k)$ and the characteristic attribute sequence $X_i(k)$ is obtained.

$$r_i = \frac{1}{N} \sum_{k=1}^{N} \xi_i(k) \tag{7}$$

The gray correlation analysis is conducted to unify all factors into a system for purposes of comparison and analysis. Therefore, it considers the correlation between all factors, which is more reasonable and scientific than the pairwise comparison method commonly used in system analysis. The

degree of closeness between the class attribute sequence curve and the feature attribute sequence curve in the system is indicated by the size of the gray relational degree, the feature attribute with the least gray relational degree has the least influence on the system's class attribute. Conversely, the feature attribute with a large gray correlation degree has a greater influence on the system category attribute. Therefore, the feature attribute with more values but less gray correlation degree has little influence on the classification result and is not the optimal attribute. Besides, considering that the curve change of sine function is relatively moderate, the correction of the information gain factor will not be excessive. Therefore, this paper introduces the sine value of the gray correlation degree as the correction factor to improve the ID3 algorithm.

4. IMPROVED ID3 ALGORITHM WITH INTELLIGENTIZED DESIGN

The specific process of the improved GBID algorithm is as Algorithm I: 1) calculates the gray correlation degree between each feature attribute and category attribute and rank them; 2) determines whether the attribute with more values is optimal through gray relational degree, so as to determine whether its information gain is reduced; 3) considers attributes with more values but lower gray correlation degrees. The sine value of the gray correlation degree is used as the correction factor when calculating their information gain, while the correction factor is set to 0 when calculating information gain for other attributes.

$$E_1(A) = \sum_{j=1}^{k} \left(\frac{s_{1j} + s_{2j} + \ldots + s_{nj}}{s} + CF(A) \right) \cdot I\left(s_{1j}, s_{2j}, \ldots, s_{nj}\right) \quad (8)$$

where $CF(A)$ is the correction factor for the attribute A, and can be expressed as:
As $0 < CF(A) < 1$ information gain for the attribute A is defined as:

$$G_1(A) = I\left(s_1, s_2, \ldots, s_n\right) - E_1(A) \quad (9)$$

5. CLASSIFYING LEARNERS BY APPLYING THE GBID ALGORITHM

The following experiments illustrate the application of the GBID algorithm. The characteristic attribute course types of all the learner test scores can be classified into A, B, C, quantified as {0,1,2}. Online learning time can be divided into short, moderate or long, quantified as {0,1,2}; the difficulty or ease of the paper is quantified as {0,1}; communication ability is strong or weak, quantified as {0,1}. The score of the classification attribute test is bounded by 80, with a score of greater than 80 being good, and bad for less than 80, is quantified as {0,1}. According to the training set of sample data, which in turn according to Eq. (6) used to calculate the gray correlation of the characteristic properties and classification, the results for r (course type) = 0.52, r (online learning time) = 0.72, r (paper difficulty) = 0.78, and r (communication ability) = 0.56; the attribute information Gain is then calculated resulting in Gain (course types) = 0.4816, Gain (online learning time) = 0.0275, Gain (paper difficulty) = 0.0588, Gain (communication ability) = 0.0368. As the information gain of the course, type is the largest, but the gray correlation degree is the lowest, the correction factor needs to be used to reduce its information gain. The correction factor CF (course type) is set as sin (0.52) = 0.4968, while the information gain of the other attributes is set to 0.

Algorithm 1 GBID (Sample_set, Attribute_set)

Input: Training sample set described by multiple attributes Sample_set; Candidate attribute set Attribute_set.
Output: A decision tree.
Begin
If Sample_set is empty
Returning null; create node **L**
If all samples in node **L** belong to the same class **C**
Then **L** is returned as the leaf node and marked with class **C**
If Attribute_set is empty;
L is returned as the leaf node, and marked with the most common class in Sample_set;
According to Eq.(4), the information gain of each attribute in Attribute_set is calculated, and the attribute **A** with the largest information gain and the attribute **B** with the largest number of values are selected;
If A=B, this condition holds that choosing the attribute with the maximum information gain and the largest number of values as the test attribute is likely to generate the multi-value bias problem, and the correction factor is needed to reduce the information gain of this attribute;
Then, calculate the correction factor of this attribute according to Eq. (8).
Then, recalculate the information gain of the attribute according to Eq. (9);
Else the correction coefficient of this attribute is 0.
The attribute with the largest information gain is not the attribute with the largest number of values.
The selection of this attribute as a split attribute will not result in the multi-value bias problem;
From Attribute_set, the Attribute with the largest information gain is selected as the Splitting_Attribute;
Tag node **L** as the Splitting_Attribute;
For the given a_i in each Splitting_Attribute ($i = 1, 2, \ldots, m$)
If Splitting_Attribute = a_i;
Corresponding branches are generated from node **L** to represent test conditions. Let S_i be the sample set obtained from Splitting_Attribute = a_i;
If S_i is empty;
Add a leaf node and mark it as the most common class in Sample_set;
Else, add the node returned by GBID (Attribute_set, Splitting_Attribute);
End

Table 1 Influence of GBID algorithm on information gain of each attribute of test score.

Characteristic attribute	Gain value of ID3 Algorithm	GBID Algorithm	
Course types	0.4816	0.4968	−0.2196
Online learning time	0.0275	0.0000	0.0275
Paper difficulty	0.0588	0.0000	0.0588
Communication ability	0.0368	0.0000	0.0368

The comparison between the GBID algorithm and the ID3 algorithm is shown in Table 1.

It can be seen from Table 1 that when the ID3 algorithm determines the root node of the decision tree, the course type with the maximum information gain is selected as the split attribute, which is obviously inconsistent with the objective fact. When determining the root node, the GBID algorithm selects the difficulty of the test paper as the split attribute, which conforms to the objective fact and avoids the course type with multi-values but not optimal attributes as the split attribute.

6. CONCLUSIONS

In the distance education system, the GBID algorithm is used to classify learners according to their test scores in course types, online learning time, paper difficulty, and communication ability, which overcomes the multi-value tendency problem of the ID3 algorithm and makes the classification more in line with objective facts. Based on this, different teaching strategies are provided for different learners in intelligentized design, to truly realize the intelligent guidance provided to each learner.

ACKNOWLEDGMENT

This work was assisted by the Hebei Xunda Public Security Laboratory.

REFERENCES

1. Wen J., Zhang W., Shu W. A. Cognitive Learning Model in Distance Education of Higher Education Institutions Based on Chaos Optimization in Big Data Environment. The Journal of Supercomputing, 2019, 75(2): 719–731.
2. El-Bishouty M. M., Aldraiweesh A., Alturki U., *et al.* Use of Felder and Silverman learning style model for online course design. Educational Technology Research and Development, 2019, 67(1): 161–177.
3. Saura J. R., Palos-Sanchez P., Grilo A. Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining. Sustainability, 2019, 11(3): 917.
4. He K. Research on Collaborative Filtering Recommendation Algorithm Based on User Interest for Cloud Computing. International Journal of Internet Manufacturing and Services, 2019, 6(4): 357–370.
5. Azzi I., Jeghal A., Radouane A., *et al.* A Robust Classification to Predict Learning Styles in Adaptive E-learning Systems. Education and Information Technologies, 2020, 25(1): 437–448.
6. Zhao J., Guo J. Online Distance Learning Precision Service Technology based on Big Data Analysis. In: 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE, 2019: 39–43.
7. Wang H., Wang T., Zhou Y., *et al.* Information Classification Algorithm based on Decision Tree Optimization. Cluster Computing, 2019, 22(3): 7559–7568.
8. Long Y. Research on Art Innovation Teaching Platform based on Data Mining Algorithm. Cluster Computing, 2019, 22(6): 14943–14949.
9. Jin D. Factors Affecting Success and Failure of Internet Company Business Model using Inductive Learning based on ID3 Algorithm. Journal of the Korea Institute of Information and Communication Engineering, 2019, 23(2): 111–116.
10. Abbas A. R., Farooq A. O. Skin Detection using Improved ID3 Algorithm. Iraqi Journal of Science, 2019: 402–410.
11. Pinto T., Morais H., Corchado J. M. Adaptive Entropy-based Learning with Dynamic Artificial Neural Network. Neurocomputing, 2019, 338: 432–440.
12. Li C., Cui P., Wang Y., *et al.* Research on Personalized Recommendation of Learning Resources Based on Data Mining. In: 2019 International Conference on Computer Network, Electronic and Automation (ICCNEA). IEEE, 2019: 62–66.
13. Sun X. Development of Intelligent English Multimedia Teaching Resources Using Data Mining. In: International Conference on Application of Intelligent Systems in Multi-modal Information Analytics. MMIA 2019. Advances in Intelligent Systems and Computing, Springer, Cham, 2019, 929: 150–155.
14. Yunlong W., Kai L., Guan G., *et al.* Evaluation Method for Green Jack-up Drilling Platform Design Scheme based on Improved Grey Correlation Analysis. Applied Ocean Research, 2019, 85: 119–127.
15. Wang J., Tang Y., Curtin A., *et al.* ECT-induced Brain Plasticity Correlates with Positive Symptom Improvement in Schizophrenia by Voxel-based Morphometry Analysis of Grey Matter. Brain stimulation, 2019, 12(2): 319–328.