An Improved ID3 Decision Tree Algorithm Based on Attribute Weighted

Xian Liang [1], Fuheng Qu[1], Yong Yang[1]
[1] College of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130022, China

Hua Cai[2]
[2] College of electronic and information engineering, Changchun University of Science and Technology, Changchun 130022, China

***Abstract*—ID3 decision tree algorithm uses information gain selection splitting attribute tend to choose the more property values, and the number of attribute values can not be used to measure the attribute importance, in view of the above problems, a new method is proposed for attribute weighting, the idea of simulation conditional probability, calculation of the close contact between the attributes and the decision attributes, as the attribute weights and combination with attribute information gain to selection splitting attribute, improve the accuracy of decision results. Experiments show that compared with the improved algorithm and the traditional ID3 algorithm, decision tree model has higher predictive accuracy, less number of leaves.**

Keywords-decision tree; ID3 algorithm; information gain

I. INTRODUCTION

ID3 algorithm[1,2,3,4] to select non leaf node in the decision tree model by introducing the definition of information gain in information theory, property has a better theoretical foundation but tend to choose the values of more, and the attribute importance from the number of attribute values cannot be measured. Aiming at the shortcomings of ID3 algorithm, The document[5]to introduce the concept of attribute importance to overcome the disadvantage of ID3 algorithm for attribute selection criteria. In document [6],the extended attributes selected with the proposed algorithm maximized the ranking mutual information between the candidate attributes and the decision attribute, and also minimized the ranking mutual information between the candidate attributes and the selected conditional arrtibutes on the same branch. The use of conditional probability calculation of close contact between the attributes and the decision attributes, as the weights and combined with information gain choose the split attribute, get the improved algorithm of ID3 decision tree algorithm based on attribute weighted.

II. AN IMPROVED ID3 DECISION TREE ALGORITHM BASED ON ATTRIBUTE WEIGHTED

Using information gain as a measure of the ID3 algorithm, without considering the relationship between classification and attribute, in the practical application have close ties between the two, the improved algorithm is proposed based on the shortcoming. The relevant definitions as: the sample data set S, the classification properties C has m different values $\{C_1, C_2, ..., C_m\}$,the data set of S is divided into m subsets S_i （i=1,2...m）, description attribute A has v different values $\{A_1, A_2, ..., A_v\}$,description attribute A set S is partitioned into v subsets S_j （j=1,2...v）, the connection between the attribute A and classification attribute C is defined as $F_A = \sum_{j=1}^{v} \frac{|S_j|}{|S|} . W_j$,The number of samples about the value of attribute A is A_j is $|S_j|$,The total number of samples is $|S|$,the sum of probability of classification results is W_j ,about the subset of samples under the condition of the value of attribute A is A_j .The way Calculation of the value of W_j as: the number of the value of attribute A is A_j is t, there are t sample $\{m_1, m_2, ..., m_t\}$, any one sample m_j （j=1,2...t）class label attribute value for C_i ,the obtained t samples under the condition of the value of attribute A is A_j ,the probability of the value of class label attribute is C_i in the t samples is $\frac{|C_{ij}|}{t}$. $|C_{ij}|$ is the number of samples in the condition of the value of the categorical attribute is C_i in the t samples. The same method can be calculated the probability in the condition of the value of classification attribute is C_i in the data sets of the value of other description attribute in the samples m_j get the corresponding sample subset, in the samples each description attribute values are independent of each other, the probability of according to the sample m_j conjecture classification attribute values of C_i ,the probability value P_j is equal to the product of the probability of each description attribute in sample m_j according to their own

values get subset of data under the condition of the value of classification attribute is C_i. under the condition of the attribute A value for B corresponding the subset of data , the sum of the probability of correct introduced classification attribute's value in the subset of data is $W_j = \sum_{j=1}^{t} P_j$ .The use of conditional probability Calculation of close contact between the attributes and the decision attributes, as the weights and combined with information gain choose the split attribute, if a property to make the $M_A = Gain(A).F_A$ value of the largest, then choose it for the splitting attribute. For example, there are a lot of value attribute id, it is the information gain is large, but it used to build classification model does not have any significance, using the method proposed in this paper to calculate the value of the relationships between Description attribute ID and classification attribute is small, so the value of $M_{\mathrm{id}} = Gain(\mathrm{id}).F_{\mathrm{id}}$ is small, avoid selecting properties such as the attribute id(the attribute value is more, but it is not important attributes) as the decision node of decision tree model.

III. THE EXPERIMENT RESULTS ANALYSIS

The data set in table 1, with establishing the decision tree use the ID3 algorithm and the improved ID3 algorithm.

TABLE 1 THE DATA SET

id	chinese	mathematics	english	physics	Summary
1	general	good	bad	general	qualified
2	general	good	good	good	qualified
3	good	general	general	good	qualified
4	optimal	general	good	good	qualified
5	general	general	general	general	qualified
6	good	bad	general	bad	unqualified
7	optimal	bad	bad	general	unqualified
8	good	optimal	optimal	optimal	qualified
9	general	general	optimal	good	qualified
10	optimal	bad	general	general	qualified
11	bad	good	good	bad	unqualified
12	good	general	good	good	qualified
13	general	bad	good	general	qualified
14	general	general	optimal	good	qualified
15	good	bad	good	general	qualified
16	optimal	general	optimal	good	qualified
17	optimal	optimal	optimal	optimal	qualified
18	good	bad	good	general	qualified
19	good	general	bad	optimal	qualified
20	general	general	general	general	qualified

ID3 algorithm build a decision tree as shown in figure 1.

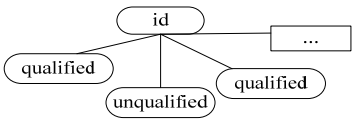

Figure 1. ID3 algorithm for constructing decision trees

Improved ID3 algorithm to construct the decision tree:
Description attribute weights respectively:
F(chinese)=3.66985.
F(mathematics)=3.9352.
F(english)=2.7904.
F(physics)=3.6665.
F(id)=0.5643.
The product of attribute weights and information gain of description attribute respectively:
M(chinese)= 0.222*3.66985=0.8147.
M(mathematics)= 0.197*3.9352=0.7752.
M(english)= 0.292*2.7904=0.8148.
M(physics)= 0.392*3.6665=1.4373.
M(id)= 0.61*0.5643=0.3443.
The improved ID3 algorithm choice of properties physical as the decision tree as shown in figure 2.

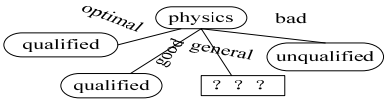

Figure 2. Improved ID3 algorithm for constructing decision subtrees 1

Improved ID3 algorithm build a decision tree as shown in figure 3.

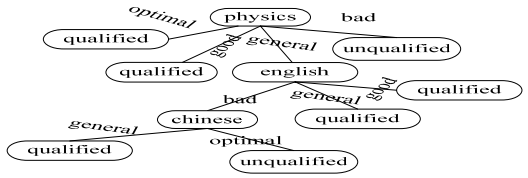

Figure 3. The decision tree built by improved ID3 algorithm

The compared to the decision tree of two kinds of algorithm, improved ID3 algorithm is not sensitive to more attribute values, build the decision tree has a lot of path, the categorization of data more detailed, to improve the decision tree classification accuracy.

In order to evaluate the performance of the improved ID3 algorithm, the experiment using seven data sets of UCI. Each data set 70% as the training data 30% as test data, the training data set is used to construct the decision tree model, the test data set is used to predict accuracy of the decision tree model. To test the traditional ID3 algorithm and the improved ID3 algorithm, the comparison results as shown in table 2.

TABLE 2 EXPERIMENTAL RESULTS COMPARING

Data sets	traditional ID3 algorithm		improved ID3 algorithm	
	Number of leaf nodes	correct （%）	Number of leaf nodes	correct（%）
Breast-cancer	93	91	84	91
Balance-scale	241	24	221	31
Car Evaluation	119	87	131	92
Nursery	372	89	359	96
Adult	386	74	375	82

The table 3 shows that, improved ID3 algorithm compared with the traditional ID3 algorithm of decision tree model accuracy is higher, the number of leaf nodes is less. Experiments show that, improved ID3 algorithm solves the ID3 algorithm tend to choose more but not important attribute values, to improve the prediction accuracy of the decision tree classification model.

IV. CONCLUSIONS

It is a project supported by the natural science foundation of Jilin Province(201215145) and the Twelfth Five-Science and Technology Research Projects of Education Department of Jilin Province (Grant No. 2013-420).The improved ID3 algorithm calculation of the close contact between the attributes and the decision attributes, as the attribute weights and combination with attribute information gain to selection splitting attribute, this method does not affect the weights of attribute is important and more attribute values is selected as split attribute, adds the weights of attributes of important but less attribute value is selected as split attribute, ensure that each time select important rather than more attribute values to construct the decision tree model, get higher prediction accuracy, less number of leaf nodes of the decision tree model.

ACKNOWLEDGEMENTS

It is a project supported by the natural science foundation of Jilin Province(20130101179JC-13) and the natural science foundation of Jilin Province(201215145) the Twelfth Five-Science and Technology Research Projects of Education Department of Jilin Province (Grant No. 2013-420). The corresponding author is Fuheng Qu.

REFERENCES

[1] Davidson Ian, Tayi Giri. Data preparation using data quality matrices for classify-cation mining. European Journal of Operational Research 2009,197(2): 764-772.

[2] Cheng Jiajun, Su Shoubao, Xu Lihua. Decision tree optimization algorithm based on multiscale rough set model, Journal of Computer Applications,2011,31(12):3243-3246.

[3] QUINLAND J R.Induction of decision trees[J],Machine Learning,1981,1(1): 81-106.

[4] Lu zhao,Cheng Shiping.Applicaion of machinery manufacturing decision-making based on ID3 algorithm[J],Journal of Computer Applications, 2011,31(11):3087-3090.

[5] Yu Jinping,Huang Ximei,Li Kangsun.Improved ID3 algorithm based on new attributes selection criterion[J],Application Research of Computers, 2012,29(8):2895-2899.

[6] Pan Pan,Wang Xizhao,Zhai Junhai.An improved induction algorithm based on ordinal decision tree[J],2014,44(1):41-44.