ISSN : 0974 - 7435

Volume 10 Issue 24

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(24), 2014 [16338-16346]

2014

Application research of decision tree algorithm in english grade analysis

Zhao Kun
Beihua University, Teacher's college, Jilin, (CHINA)

ABSTRACT

This paper introduces and analyses the data mining in the management of students' grades. We use the decision tree in analysis of grades and investigate attribute selection measure including data cleaning. We take course score of institute of English language for example and produce decision tree using ID3 algorithm which gives the detailed calculation process. Because the original algorithm lacks termination condition, we propose an improved algorithm which can help us to find the latency factor which impacts the grades.

KEYWORDS

Decision tree algorithm; English grade analysis; ID3 algorithm; Classification.

INTRODUCTION

With the rapid development of higher education, English grade analysis as an important guarantee for the scientific management constitutes the main part of the English educational assessment. The research on application of data mining in management of students' grades wants to talk how to get the useful uncovered information from the large amounts of data with the data mining and grade management[1-5]. It introduces and analyses the data mining in the management of students' grades. It uses the decision tree in analysis of grades. It describes the function, status and deficiency of the management of students' grades. It tells us how to employ the decision tree in management of students' grades. It improves the ID3 arithmetic to analyze the students' grades so that we could find the latency factor which impacts the grades. If we find out the factors, we can offer the decision-making information to teachers. It also advances the quality of teaching[6-10]. The English grade analysis helps teachers to improve the teaching quality and provides decisions for school leaders.

The decision tree-based classification model is widely used as its unique advantage. Firstly, the structure of the decision tree method is simple and it generates rules easy to understand. Secondly, the high efficiency of the decision tree model is more appropriate for the case of a large amount of data in the training set. Furthermore the computation of the decision tree algorithm is relatively not large. The decision tree method usually does not require knowledge of the training data, and specializes in the treatment of non-numeric data. Finally, the decision tree method has high classification accuracy, and it is to identify common characteristics of library objects, and classify them in accordance with the classification model. The original decision tree algorithm uses the top-down recursive way[11-12]. Comparison of property values is done in the internal nodes of the decision tree and according to the different property values judge down branches from the node. We get conclusion from the decision tree leaf node. Therefore, a path from the root to the leaf node corresponds to a conjunctive rules, the entire decision tree corresponds to a set of disjunctive expressions rules. The decision tree generation algorithm is divided into two steps[13-15]. The first step is the generation of the tree, and at the beginning all the data is in the root node, then do the recursive data slice. Tree pruning is to remove some of the noise or abnormal data. Conditions of decision tree to stop splitting is that a node data belongs to the same category and there are not attributes used to split the data.

In the next section, we introduce construction of decision tree. In Section 3 we introduce attribute selection measure. In Section 4, we do empirical research based on ID3 algorithm and propose an improved algorithm. In Section 5 we conclude the paper and give some remarks.

CONSTRUCTION OF DECISION TREE USING ID3

The growing step of the decision tree is shown in Figure 1. Decision tree generation algorithm is described as follows. The name of the algorithm is $Generate\_decision\_tree$ which produce a decision tree by given training data. The input is training samples which is represented with discrete values. Candidate attribute set is attribute. The output is a decision tree.

Step1. Set up node N. If samples is in a same class C then return N as lead node and label it with C.

Step2. If attribute_list is empty, then return N as leaf node and label it with the most common class in the samples.

Step3. Choose $test\_attribute$ with information gain in the attribute_list, and label N as $test\_attribute$.

Step4. While each a_i in every $test\_attribute$ do the following operation.

Step5. Node N produces a branch which meets the condition of $test\_attribute = a_i$

Step6. Suppose s_i is sample set of $test\_attribute = a_i$ in the samples. If s_i is empty, then plus a leaf and label it as the most common class. Otherwise plus a node which was returned by $Generate\_decision\_tree(s_i, attribute\_list - test\_attribute)$.

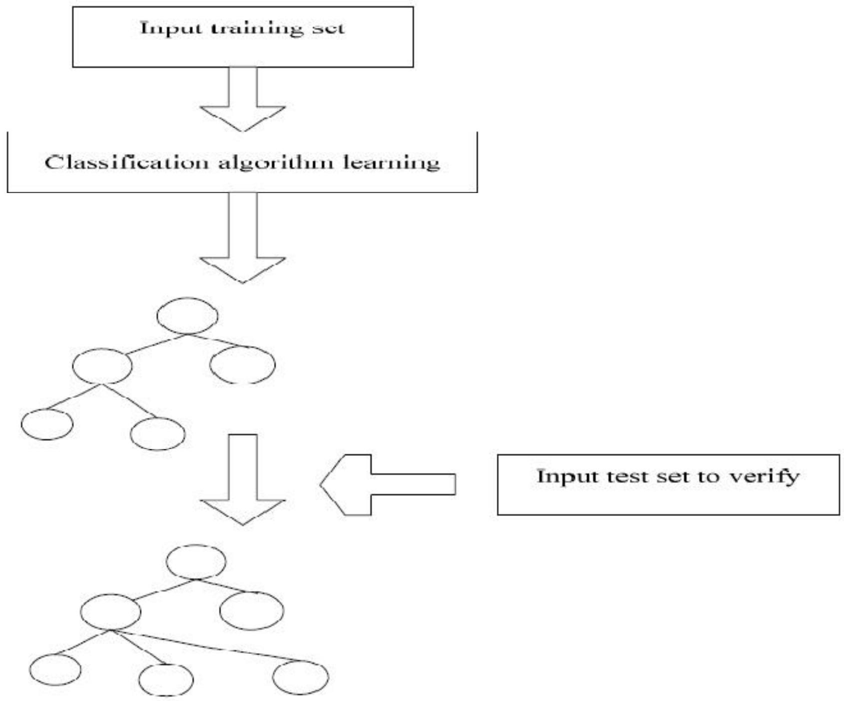

Figure 1: Growing step of the decision tree

AN IMPROVED ALGORITHM

Attribute selection measure

Suppose S is data sample set of s number and class label attribute has m different values $C_i(i=1,2,\cdots,m)$. Suppose S_i is the number of sample of class C_i in S. For a given sample classification the demanded expectation information is given by formula 1[11-12].

$$I(s_{1j},s_{2j},K,s_{mj}) = -\sum_{i=1}^{m} p_{ij}\log_2 p_{ij}(i=1,2,\cdots,K,n) \tag{1}$$

$$E(A)=\sum_{j=1}^{V}\frac{(S_{1j}+S_{2j}+\cdots+S_{mj})}{S}I(S_{1j},S_{2j},K,S_{mj}) \tag{2}$$

p_i is probability that random sample belongs to C_i and is estimated by s_i/s. Suppose attribute A has V different values $(a_1,a_2,\cdots,a_V)$. We can use attribute A to classify S into V number of subset $(S_1,S_2,\cdots S_V)$. Suppose S_{ij} is the number of class C_i in subset S_j. The expected information of subset is shown in formula 2. $\frac{(S_{1j}+S_{2j}+\cdots+S_{mj})}{S}$ is the weight of the j-th subset. For a given subset S_j formula 3 sets up[13].

$$I(s_{1j},s_{2j},K,s_{mj}) = -\sum_{i=1}^{m} p_{ij}\log_2 p_{ij}(i=1,2,\cdots,K,n) \tag{3}$$

$p_{ij} = \frac{s_{ij}}{|s_j|}$ is the probability that samples of s_j belongs to class C_i. If we branch in A, the information gain is shown in formula 4[14].

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \quad \textbf{(4)}$$

The improved algorithm

The improved algorithm is as follows. Function $Generate\_decision\_tree$ (training samples, candidate attribute attribute_list)

{Set up node N;
If samples are in the same class C then
Return N as leaf node and label it with C;
Record statistical data meeting the conditions on the leaf node;
If attribute_list is empty then
Return N as the leaf node and label it as the most common class of samples;
Record statistical data meeting the conditions on the leaf node;
Suppose GainMax=max (Gain1, Gain2, …, Gainn)
If GainMax< threshold
Return N as the leaf node and label it as the most common class of samples;
Choose attribute with the highest information gain of attribute_list;
Label N as test_attribute;
For each a_i of test_attribute, produce a branch from node N meeting the condition of test_attribute= a_i;
Suppose s_i sample set of samples meeting the condition of test_attribute= a_i;
If s_i is empty then Record statistical data meeting the conditions on the leaf node;Add a leaf and label it as the most common class of samples;
Else add a node returned by $Generate\_decision\_tree$ (s_i, attribute_list_test_attribute);
}

EMPIRICAL RESEARCH

Data cleaning

This paper takes course score of institute of English language for example. Examination score of the students is shown in TABLE 1.

TABLE 1: Examination score of the students

Course code	Whether re-learning	Paper difficulty	Whether required course	score
110101290	no	high	yes	89
H200104088	no	middle	yes	75
H2001 16090	yes	middle	no	80
H120101160	yes	high	yes	65
120101288	yes	middle	yes	70
H200152069	no	low	no	90

Data in TABLE 1 is not suitable for classification, so we firstly do data cleaning. According to the general course, basic course, professional basic course and specialized course, classify the course into A, B, C, D. Score is divided into three categories outstanding, medium, general. Paper difficulty is is divided into three categories 1, 2, 3. Such as

Update ks set ci_pi='outstanding' where ci_pj>='85'
Update ks set ci_pi=' medium' where ci_pj>='75'and ci_pj<'85'
Update ks set ci_pi='general' where ci_pj>='60'and ci_pj<'75'
Update ks set sjnd='high' where sjnd='1'
Update ks set sjnd=' medium' where sjnd='2'
Update ks set sjnd='low' where sjnd='3'

Result of ID3 algorithm

TABLE 2 is training set of student test scores situation information after data cleaning. We classify the samples into three categories. $C_1 = "\text{outstanding}"$, $C_2 = "\text{medium}"$, $C_3 = "\text{general}"$, $s_1 = 300$, $s_2 = 1950$, $s_3 = 880$, $s = 3130$. According to formula 1, we obtain $I(s_1, s_2, s_3) = (300, 1950, 880) = -(300/3130)\ /\log_2(300/3130)$.

$$-(1950/3130)\log_2(1950/3130) - (880/3130)\log_2(880/3130)$$
$$= 1.256003.$$

Entropy of every attribute is calculated as follows. Firstly calculate whether re-learning. For yes, $s_{11} = 210$, $s_{21} = 950$, $s_{31} = 580$.

$$I(s_{11}, s_{21}, s_{31}) = (210, 950, 580)$$
$$= -(210/1740)\log_2(210/1740) - (950/1740)\log_2(950/1740)$$
$$-(580/1740)\log_2(580/1740) = 1.074901$$

For no, $s_{12} = 90$, $s_{22} = 1000$, $s_{32} = 300$.

$$I(s_{12}, s_{22}, s_{32}) = (90, 1000, 300)$$
$$= -(90/1390)\log_2(90/1390) - (1000/1390)\log_2(1000/1390)$$
$$-(300/1390)\log_2(300/1390) = 1.373186.$$

IF samples are classified according to whether re-learning, the expected information is

$$E("whether \quad re\text{-}learning") = (1740/3130)\cdot I(s_{11}, s_{21}, s_{31}) + (1390/3130)\cdot I(s_{12}, s_{22}, s_{32})$$

$= 0.555911 \cdot 1.074901 + 0.444089 \cdot 1.373186 = 1.240721$. So the information gain is

$$Gain("whether \quad re\text{-}learning") = I(s_1, s_2, s_3) - E("whether \quad re\text{-}learning") = 0.015282.$$

Secondly calculate course type, when it is A, $s_{11} = 110, s_{21} = 200, s_{31} = 580$.

$$I(s_{11}, s_{21}, s_{31}) = (110, 200, 580)$$
$$= -(110/890)\log_2(110/890) - (200/890)\log_2(200/890) - (580/890)\log_2(580/890)$$
$$= 1.259382.$$

For course type B, $s_{12} = 100, s_{22} = 400, s_{32} = 0$.

$$I(s_{12}, s_{22}, s_{32}) = (100, 400, 0)$$
$$= -(100/500)\log_2(100/500) - (400/500)\log_2(400/500) - 0$$
$$= 0.721928\,.$$

For course type C, $s_{13} = 0, s_{23} = 550, s_{33} = 0$.

$$I(s_{13}, s_{23}, s_{33}) = (0, 550, 0)$$
$$= -(0/550)\log_2(0/550) - (550/500)\log_2(550/500) - 0$$
$$= 1.168009\,.$$

For course type D, $s_{14} = 90, s_{24} = 800, s_{34} = 300$.

$$I(s_{14}, s_{24}, s_{34}) = (90, 800, 300)$$
$$= -(90/1190)\log_2(90/1190) - (800/1190)\log_2(800/1190) - (300/1190)\log_2(300/1190) = 1.168009\,.$$

$$E(\text{"}course \quad type\text{"}) = (890/3130)\cdot I(s_{11}, s_{21}, s_{31}) + (500/3130)\cdot I(s_{12}, s_{22}, s_{32})$$

$$+(550/3130)\cdot I(s_{13}, s_{23}, s_{33}) + (1190/3130)\cdot I(s_{14}, s_{24}, s_{34})$$
$$= 0.91749.$$

$$Gain(\text{"}course \quad type\text{"}) = 1.256003 - 0.91749 = 0.338513\,.$$

Thirdly calculate paper difficulty. For high, $s_{11} = 110, s_{21} = 900, s_{31} = 280$.

$$I(s_{11}, s_{21}, s_{31}) = (110, 900, 280)$$
$$= -(110/1290)\log_2(110/1290) - (900/1290)\log_2(900/1290) - (280/1290)\log_2(280/1290) = 1.14385\,.$$

For medium, $s_{12} = 190, s_{22} = 700, s_{32} = 300$.

$$I(s_{12}, s_{22}, s_{32}) = (190, 700, 300)$$
$$= -(190/1190)\log_2(190/1190) - (700/1190)\log_2(700/1190) - (300/1190)\log_2(300/1190) = 1.374086$$

For low, $s_{13} = 0, s_{23} = 350, s_{33} = 300$.

$$I(s_{13}, s_{23}, s_{33}) = (0, 350, 300)$$
$$= -(0/650)\log_2(0/650) - (350/650)\log_2(350/650) - (300/650)\log_2(300/650) = 0.995727.$$

$$E(\text{"}paper \quad difficulty\text{"}) = (1290/3130)\cdot I(s_{11}, s_{21}, s_{31}) + (1190/3130)\cdot I(s_{12}, s_{22}, s_{32})$$

$$+(650/3130)\cdot I(s_{13}, s_{23}, s_{33}) = 1.200512.$$

$$Gain(\text{"}paper \quad difficulty\text{"}) = 1.256003 - 1.200512 = 0.55497.$$

Fourthly calculate whether required course. For yes, $s_{11} = 210, s_{21} = 850, s_{31} = 600$

$$I(s_{11}, s_{21}, s_{31}) = (210, 850, 600)$$
$$= -(210/1660)\log_2(210/1660) - (850/1660)\log_2(850/1660) - (600/1660)\log_2(600/1660) = 1.220681.$$

For no, $s_{12} = 90, s_{22} = 1100, s_{32} = 280$

$$I(s_{12}, s_{22}, s_{32}) = (90, 1100, 280)$$
$$= -(90/1470)\log_2(90/1470) - (1100/1470)\log_2(1100/1470) - (280/1470)\log_2(280/1470) = 1.015442.$$
$$E(\text{"}whether \quad required\text{"}) = (1660/3130)\cdot I(s_{11}, s_{21}, s_{31}) + (1470/3130)\cdot I(s_{12}, s_{22}, s_{32})$$
$$= 1.220681.$$

$$Gain(\text{"}whether \quad required\text{"}) = 1.256003 - 1.220681 = 0.035322.$$

TABLE 2: Training set of student test scores

Course type	Whether re-learning	Paper difficulty	Whether required	Score	Statistical data
D	no	medium	no	outstanding	90
B	yes	medium	yes	outstanding	100
A	yes	high	yes	medium	200
D	no	low	no	medium	350
C	yes	medium	yes	general	300
A	yes	high	no	medium	250
B	no	high	no	medium	300
A	yes	high	yes	outstanding	110
D	yes	medium	yes	medium	500
D	no	low	yes	general	300
A	yes	high	no	general	280
B	no	high	yes	medium	150
C	no	medium	no	medium	200

Result of improved algorithm

The original algorithm lacks termination condition. There are only two records for a subtree to be classified which is shown in TABLE 3.

TABLE 3: Special case for classification of the subtree

Course type	Whether re-learning	Paper difficulty	Whether required	Score	Statistical data
A	no	high	yes	medium	15
A	no	high	yes	general	20

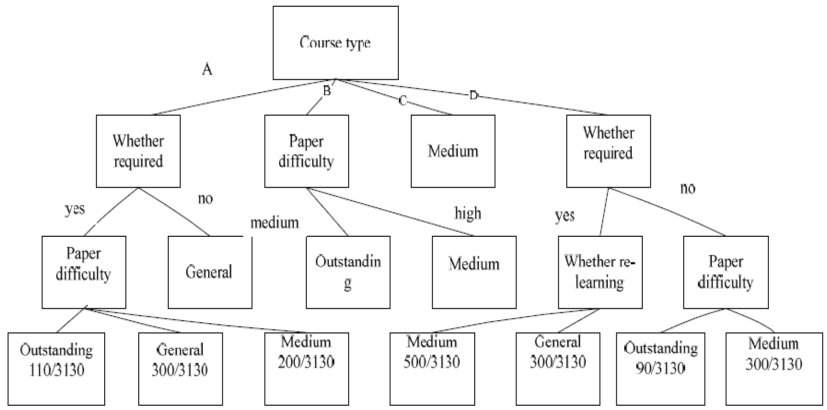

Figure 2: Decision tree using improved algorithm

All Gains calculated are 0.00, and GainMax=0.00 which does not conform to recursive termination condition of the original algorithm in TABLE 3. The tree obtained is not reasonable, so we adopt the improved algorithm and decision tree using improved algorithm is shown in Figure 2.

CONCLUSIONS

In this paper we study construction of decision tree and attribute selection measure. Because the original algorithm lacks termination condition, we propose an improved algorithm. We take course score of institute of English language for example and we could find the latency factor which impacts the grades.

REFERENCES

[1] Xuelei Xu, Chunwei Lou; "Applying Decision Tree Algorithms in English Vocabulary Test Item Selection", IJACT: International Journal of Advancements in Computing Technology, **4(4)**, 165-173 **(2012)**.

[2] Huawei Zhang; "Lazy Decision Tree Method for Distributed Privacy Preserving Data Mining", IJACT: International Journal of Advancements in Computing Technology, **4(14)**, 458-465 **(2012)**.

[3] Xin-hua Zhu, Jin-ling Zhang, Jiang-tao Lu; "An Education Decision Support System Based on Data Mining Technology", JDCTA: International Journal of Digital Content Technology and its Applications, **6(23)**, 354-363 **(2012)**.

[4] Zhen Liu, Xian Feng Yang; "An application model of fuzzy clustering analysis and decision tree algorithms in building web mining", JDCTA: International Journal of Digital Content Technology and its Applications, **6(23)**, 492-500 **(2012)**.

[5] Guang-xian Ji; "The research of decision tree learning algorithm in technology of data mining classification", JCIT: Journal of Convergence Information Technology, **7(10)**, 216-223 **(2012)**.

[6] Fuxian Huang; "Research of an Algorithm for Generating Cost-Sensitive Decision Tree Based on Attribute Significance", JDCTA: International Journal of Digital Content Technology and its Applications, **6(12)**, 308-316 **(2012)**.

[7] M.Sudheep Elayidom, Sumam Mary Idikkula, Joseph Alexander; "Design and Performance analysis of Data mining techniques Based on Decision trees and Naive Bayes classifier For", JCIT: Journal of Convergence Information Technology, **6(5)**, 89-98 **(2011)**.

[8] Marjan Bahrololum, Elham Salahi, Mahmoud Khaleghi; "An Improved Intrusion Detection Technique based on two Strategies Using Decision Tree and Neural Network", JCIT: Journal of Convergence Information Technology, **4(4)**, 96-101 **(2009)**.

[9] Bor-tyng Wang, Tian-Wei Sheu, Jung-Chin Liang, Jian-Wei Tzeng, Nagai Masatake; "The Study of Soft Computing on the Field of English Education: Applying Grey S-P Chart in English Writing Assessment", JDCTA: International Journal of Digital Content Technology and its Applications, **5(9)**, 379-388 **(2011)**.

[10] Mohamad Farhan Mohamad Mohsin, Mohd Helmy Abd Wahab, Mohd Fairuz Zaiyadi, Cik Fazilah Hibadullah; "An Investigation into Influence Factor of Student Programming Grade Using Association Rule Mining", AISS: Advances in Information Sciences and Service Sciences, **2(2)**, 19-27 **(2010)**.

[11] Hao Xin; "Assessment and Analysis of Hierarchical and Progressive Bilingual English Education Based on Neuro-Fuzzy approach", AISS: Advances in Information Sciences and Service Sciences, **5(1)**, 269-276 **(2013)**.

[12] Hong-chao Chen, Jin-ling Zhang, Ya-qiong Deng; "Application of Mixed-Weighted-Association-Rules-Based Data Mining Technology in College Examination grades Analysis ", JDCTA: International Journal of Digital Content Technology and its Applications, **6(10)**, 336-344 **(2012)**.

[13] Yuan Wang, Lan Zheng; "Endocrine Hormones Association Rules Mining Based on Improved Apriori Algorithm", JCIT: Journal of Convergence Information Technology, **7(7)**, 72-82 **(2012)**.

[14] Tian Bai, Jinchao Ji, Zhe Wang, Chunguang Zhou; "Application of a Global Categorical Data Clustering Method in Medical Data Analysis", AISS: Advances in Information Sciences and Service Sciences, **4(7)**, 182-190 **(2012)**.

[15] HongYan Mei, Yan Wang, Jun Zhou; "Decision Rules Extraction Based on Necessary and Sufficient Strength and Classification Algorithm", AISS: Advances in Information Sciences and Service Sciences, **4(14)**, 441-449 **(2012)**.

[16] Liu Yong; "The Building of Data Mining Systems based on Transaction Data Mining Language using Java", JDCTA: International Journal of Digital Content Technology and its Applications, **6(14)**, 298-305 **(2012)**.