International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.31-38
http://dx.doi.org/10.14257/ijdta.2016.9.8.04

Decision Making Based on Data Mining for Traditional Sports Personnel Training Scheme

Hu Xiaoyong, Yin Yujia, Wang Yuqian and Kang Le

Guiyang University, Guiyang Guizhou, 550005 China
lanqiu727@163.com

Abstract

The percent of examinees' first applications for a university can reflect the scientific research level of this university, so various universities start to research how to improve the percent of examinees' first applications under the condition of not influencing the enrollment quality. For such research, C4.5 decision tree algorithm is applied to the postgraduate enrollment of a certain university. Specifically, examinees' information is processed to select decision attributes and establish the decision tree so as to obtain the relation among examinees' first applications, native place information, total points of initial examination and category of graduation universities from the rules extracted thereby. The mining result shows that this algorithm can correctly classify the graduation universities and assist the enrolling personnel to more effectively stipulate the enrollment guide for the targeted enrollment propaganda, thus to improve the percent of examinees' first applications.

Keywords*: Algorithm; Decision tree; Enrollment; Percent of examinees' first applications*

1. Introduction

Along with the rapid expansion of the postgraduate education scale in China, the postgraduate enrollment quality of the sports major is concerned by various universities all the time. Therefore, various universities continuously take many measures for enrollment competition for many years. During the competition process, due to the limitation upon school running conditions, subject setting, doctor program availability and other factors, some universities may have low percent of examinees' first applications. Percent of examinees' first applications can not only represent the popularity of a university, but also more or less represent the scientific research level of this university. Therefore, how to improve the percent of examinees' first applications under the condition of ensuring the enrollment quality becomes a core issue concerned by some universities. As an expression method for the decision-making process of judging the correlation between the given sample and a certain attribute, the decision tree is mainly used to solve relevant classification problems. At present, the decision tree generation algorithms include CART algorithm, ID3 algorithm, C4.5 algorithm, *etc.*, wherein C4.5 algorithm is widely applied due to the features of fast classification speed and high accuracy.

On the basis of combining data mining technology in this paper, C4.5 decision tree algorithm is selected for mining and analyzing the data regarding the adjusted examinees of university M. The analysis result can assist university M to clearly find propaganda emphasis and effectively carry out relevant propaganda work, thus to improve the percent of examinees' first applications for this university.

2. C4.5 Algorithm Introduction

As a relatively perfect decision tree classification algorithm improved on the basis of the famous classification algorithm ID3, C4.5 algorithm is a classic decision tree algorithm. C4.5 algorithm is improved from the following aspects: (1) the processing continuity attribute is improved and this is a key improvement; (2) the information gain is replaced by the information gain ratio as the standard for branch attribute selection in order to solve the problem of multiple deviation selections; (3) incomplete data can be also processed; (4) pruning operation is executed during the tree establishment process [1-4].

2.1. C4.5 Algorithm

C4.5 algorithm mainly includes the following processing steps:

(1) Cluster information entropy

Set S is set as a set of s training samples and totally includes m cluster samples Ci (i=1,2,…,m), and si is the number of the samples of Ci cluster in set S. For a given sample, the expected information needed for classification is calculated according to Formula (1) [5-9]:

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{1}$$

Wherein pi is the probability for the sample to belong to Ci and can be estimated by si/s.

(2) Cluster condition entropy

Attribute A is assumed to have v different values {a1,a2,…,av}, and can be adopted to divide set S into v subsets {S1, S2,…, Sv}, wherein set Sj is a subset of set S and the samples in Sj have the same value aj (j=1,2,…,v) for attribute A. The condition entropy needed for the classification based on attribute A can be calculated according to Formula (2) [5-9]:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + ... + s_{mj}}{s} I(s_{1j}, ...s_{mj}) \tag{2}$$

Wherein sij is the number of the samples belonging to cluster Ci in subset Sj, and I(s1j,…,smj) is set as I(s1j,…,smj)= $-\sum_{i=1}^{m} P_{ij} \log_2(P_{ij})$.

$\mathrm{P}_{ij} = \frac{s_{ij}}{|s_j|}$ is the probability for the samples in subset Sj to belong to cluster Ci.

(3) Information gain

The information gain function of attribute A is as shown in Formula (3) [5-9]:

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \tag{3}$$

(4) Information gain ratio

The information gain ratio function is as shown in Formula (4) [5-9]:

$$Gainratio(A) = \frac{Gain(A)}{I(s_1, s_2, ..., s_v)} \tag{4}$$

Wherein v is the number of the branches of the node, and si the number of the records under the *i*th branch.

2.2. Tree Pruning

Abnormal branches may be generated due to data noise, *etc.* during the decision tree establishment process, so tree pruning method is proposed in order to eliminate the above phenomena. The thought of tree pruning method is to adopt the statistical measures to find and prune the most unreliable branch. The pruned decision tree can improve the capability for correct data classification. Common tree pruning methods includes prepruning method and postpruning method, wherein the prepruning method means that the decision tree establishment process is suspended in advance in order to prune the unreliable branches, and once the prepruning method is terminated, the node concern will become a leaf; the postpruning method means that the abnormal branches are pruned from a completely established decision tree [5, 10-12].

3. Decision Tree Establishment of C4.5 Algorithm

3.1. Data Preprocessing

The enrollment and admission data of the sports major of university M in the freshmen year are adopted in this paper. Based on previous research object, the examinees data of A college involving in much enrollment adjustment in this university are taken as the training dataset for rule extraction. After data preprocessing, the following five attributes, namely: examinees' graduation universities, examinees' first applications, province/city codes of examinees' native places, sources of examinees and total points of initial examination, are adopted to classify and process the examinees data, and the above attributes are respectively numbered as a1, a2, a3, a4 and a5.

The application information of the examinees is collected by computers, so the data are relatively accurate and the data preprocessing shall be mainly focused on attribute value discretization [11]. The specific attribute discretization results are as follows:

Graduation universities (a1): the graduation universities are discretized into the universities affiliated to the Ministry of Education, 985 universities, 211 universities and other institutions according to the school subordination relationship and the school running conditions;

Examinees' first applications (a2): examinees' first applications are discretized into the universities affiliated to the Ministry of Education, 985 universities, 211 universities and other institutions according to the school subordination relationship and the school running conditions;

Province/city codes of examinees' native places (a3): examinees' native places are discretized into North China, Northeast China, East China, Central China, South China, West China and special regions according to the regional divisions in China;

Sources of examinees (a4): this attribute is discretized into fresh graduates and the graduates for less than three years, for less than five years and for more than five years according to the graduation time of these examinees;

Total points of initial examination (a5): the total points are firstly converted into standard scores according to Formula (5) and then classified into grades A, B, C, D and E according to the above standard scores.

$$Z\_grade = \frac{(Total\ Points\ of\ Initial\ Examination - Mean\ Value\ of\ Total\ Points\ of\ Initial\ Examination)}{Standard\ Deviation\ of\ Examinees'\ Total\ Points\ of\ Initial\ Examination} \quad (5)$$

After attribute discretization, the repeated objects shall be deleted to finish the data preprocessing.

According to the research subject of this paper, the attribute ---- examinees' graduation universities (a1) is taken as the classification attribute, and other attributes ---- examinees' first applications (a2), province/city codes of examinees' native places (a3), sources of examinees (a4) and total points of initial examination (a5) are taken as the decision attributes[13].

3.2. Establishment of Decision Tree Model

After data preprocessing, C4.5 algorithm introduced in this paper is adopted to generate the decision tree model for the graduation universities (a1) of the examinees of the sports major. Firstly, the information entropy needed for the classification of the given samples is calculated according to Formula (1);

$$I(3,2,1,5,46)=$$
$$-\frac{3}{57}\log_2\frac{3}{57}-\frac{2}{57}\log_2\frac{2}{57}-\frac{1}{57}\log_2\frac{1}{57}$$
$$-\frac{5}{57}\log_2\frac{5}{57}-\frac{46}{57}\log_2\frac{46}{57}=1.0530939$$

Then, the information gain ratio of each decision attribute is calculated, wherein the decision attribute ---- examinees' first applications (a2) is taken as an example to calculate the expected information entropies under the four categories, namely 211 universities, 985 universities, other institutions and universities affiliated to the Ministry of Education, thus to obtain the information gain ratio of this attribute.

(1) When a2 is a 211 university, the following value is obtained according to Formula (3):

$$I(0,0,0,0,1)=-\frac{1}{1}\log_2\frac{1}{1}=0$$

(2) When a2 is a 985 university, the following value is obtained according to Formula (3):

$$I(0,1,0,1,10)=-\frac{1}{12}\log_2\frac{1}{12}-\frac{1}{12}\log_2\frac{1}{12}$$
$$-\frac{10}{12}\log_2\frac{10}{12}=0.8166891$$

(3) When a2 is any other institution, the following value is obtained according to Formula (3):

$$I(0,0,0,0,0)=0\ ;$$

(4) When a2 is a university affiliated to the Ministry of Education, the following value is obtained according to Formula (3):

$$I(3,1,1,4,35)=-\frac{3}{44}\log_2\frac{3}{44}-\frac{1}{44}\log_2\frac{1}{44}-$$
$$\frac{1}{44}\log_2\frac{1}{44}-\frac{4}{44}\log_2\frac{4}{44}-\frac{35}{44}\log_2\frac{35}{44}$$
$$=1.0894363$$

Therefore, the expected information entropy of decision attribute a2 is calculated as follows according to Formula (2):

$$E(\mathrm{a2})=\frac{1}{57}I(0,0,0,0,1)+\frac{12}{57}I(0,1,0,1,10)+$$
$$\frac{0}{57}I(0,0,0,0,0)+\frac{44}{57}I(3,1,1,4,35)$$
$$=1.0129029$$

The information gain of decision attribute a2 is calculated as follows according to Formula (3):

$$Gain(\mathrm{a2})=I(3,2,1,5,46)-E(\mathrm{a2})$$
$$=0.040191$$
$$Spliti(\mathrm{a2})=-\frac{1}{57}\times\log 2\frac{1}{57}$$
$$-\frac{12}{57}\times\log 2\frac{12}{57}-\frac{44}{57}\times\log 2\frac{44}{57}$$
$$=0.86386$$

The information gain ratio of decision attribute a2 is calculated as follows according to Formula (4)

$$Gainratio(\mathrm{a2})=\frac{Gain(\mathrm{a2})}{Spliti(\mathrm{a2})}=0.0465$$

According to Formulae (1)~(5), the information gain ratios of the decision attributes ---- province/city codes of examinees' native places (a3), sources of examinees (a4) and total points of initial examination (a5) are respectively 0.1477, 0.0999 and 0.1249.

Obviously, decision attribute 3a has the maximum information gain ratio, so this attribute is taken as the root node of the decision tree and each attribute value is introduced with one branch. Each branch node is further divided according to Formulae (1)~(5) till the decision tree is completely established.

After decision tree establishment, the postpruning method is adopted in this paper to avoid over-training, and the pruned decision tree is as shown in Figure 1.

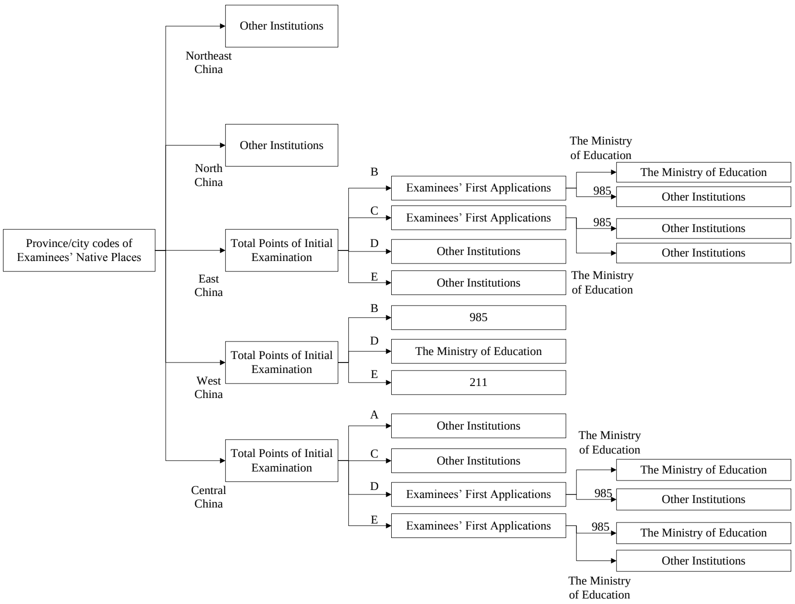
Province/city codes of Examinees' Native Places
Other Institutions
Northeast China
Other Institutions
North China
Total Points of Initial Examination
East China
B
Examinees' First Applications
The Ministry of Education
The Ministry of Education
985
Other Institutions
C
Examinees' First Applications
985
Other Institutions
Other Institutions
The Ministry of Education
D
Other Institutions
E
Other Institutions
Total Points of Initial Examination
West China
B
985
D
The Ministry of Education
E
211
Total Points of Initial Examination
Central China
A
Other Institutions
C
Other Institutions
The Ministry of Education
The Ministry of Education
D
Examinees' First Applications
985
Other Institutions
E
Examinees' First Applications
985
The Ministry of Education
Other Institutions
The Ministry of Education

Figure 1. Decision Tree for Examinees' Graduation Universities

3.3. Classification Generation Rule

The most significant advantage of the decision tree lies in the direct extraction of the classification rule [14]. Due to the paper length limitation, only some rules are listed as follows:

IF Province/City Codes of Examinees' Native Place = 'East China' AND Total Points of Initial Examination = 'B' ANDE Examinees' First Applications = '985' THEN Graduation Universities= 'Other Institutions' 50%

IF Province/City Codes of Examinees' Native Place = 'East China' AND Total Points of Initial Examination = 'C' ANDE Examinees' First Applications = 'the Ministry of Education' THEN Graduation Universities= 'Other Institutions' 50%

IF Province/City Codes of Examinees' Native Place = 'East China' AND Total Points of Initial Examination = 'D' THEN Graduation Universities= 'Other Institutions' 90%

IF Province/City Codes of Examinees' Native Place = 'East China' AND Total Points of Initial Examination = 'E' THEN Graduation Universities= 'Other Institutions' 83.3%

IF Province/City Codes of Examinees' Native Place = 'West China' AND Total Points of Initial Examination = 'B' THEN Graduation Universities= '985' 33.3%

IF Province/City Codes of Examinees' Native Place = 'West China' AND Total Points of Initial Examination = 'D' THEN Graduation Universities= 'the Ministry of Education' 33.3%

IF Province/City Codes of Examinees' Native Place = 'West China' AND Total Points of Initial Examination = 'E' THEN Graduation Universities= '211' 33.3%

IF Province/City Codes of Examinees' Native Place = 'Central China' AND Total Points of Initial Examination = 'D' ANDE Examinees' First Applications = 'the Ministry of Education' THEN Graduation Universities= 'Other Institutions' 60%

IF Province/City Codes of Examinees' Native Place = 'Central China' AND Total Points of Initial Examination = 'E' ANDE Examinees' First Applications = 'the Ministry of Education' THEN a1= 'the Ministry of Education' 50%

3.4. Result Analysis

According to the decision tree as shown in Figure 1 and the rules set generated thereby, the following conclusions can be obtained: firstly, most of the universities applied by the adjusted examinees of this college are the universities affiliated to the Ministry of Education and 985 universities, but most adjusted examinees are graduated from common graduation universities. Secondly, attribute a 3 ---- province/city codes of examinees' native places has strong dependence relation with attribute a1 ----- graduation universities; college A mainly enrolls the examinees from East China, North China and Central China when selecting the adjusted examinees. Thirdly, it can be found in the decision tree that the examinees from Central China and West China are also excellent; although the examinees are from the regions less prosperous than East China, yet the graduation universities thereof are also at a high level.

4. Conclusion

According to above research and analysis, the following reference policy suggestions can be obtained:

(1) Strengthen propaganda and provide preferential policy for the excellent students of sports major: it can be found in the data mining results that the universities firstly applied by the examinees adjusted to this college are the universities affiliated to the Ministry of Education and 985 universities but these examinees are graduated from common universities, thus indicating that the excellent students in common universities shall be taken as the key propaganda objects of university M. The sports universities shall strengthen the propaganda in the common universities with relatively high school running conditions, make full use of network resources to timely advertise their subject and major advantages and the enrollment policy to the society. For example, university M can provide such preferential policy as summer school or recommendation qualification to attract the excellent students of common universities to take university M as their first applications.

(2) Pay attention to the examinees from West China and encourage them to apply for this university: although the geographical conditions of these regions are inferior to those of coastal regions, yet the graduation universities of these examinees are also at a relatively high level. Therefore, it is necessary to positively advertise to the students and the teachers in West China and meanwhile follow the national policy for western development.

(3) Pay attention to regional factors during the enrollment process: it can be found in the data mining results that many examinees may consider the issues regarding future employment and parents care during the university application process, so it is necessary to make targeted regional advantage propaganda according to the sources of the examinees so as to attract examinees to apply for this university.

In this paper, the decision tree algorithm is adopted for the data mining of the postgraduate enrollment database of the sports major. Some significant results obtained thereby are favorable for improving the percent of examinees' first applications for this university on the basis of ensuring the student quality. However, due to time and space limitations, these data have limited guiding function for the enrollment propaganda decision. Therefore, the data quality shall be improved in order to ensure the accuracy of the classification rules, thus to provide better reference basis for the postgraduate enrollment work of university M.

Acknowledgement

The research is supported by Guizhou Provincial Education Department of Humanities and social sciences research project(14zc285).

References

[1] T. Su, W. Wang and Z. Lv, "Rapid Delaunay triangulation for randomly distributed point cloud data using adaptive Hilbert curve", Computers & Graphics, vol. 54, **(2016)**, pp. 65-74.

[2] N. Lu, C. Lu, Z. Yang and Y. Geng, "Modeling Framework for Mining Lifecycle Management", Journal of Networks, vol. 9, no. 3, **(2014)**, pp. 719-725.

[3] Y. Geng and K. Pahlavan, "On the accuracy of rf and image processing based hybrid localization for wireless capsule endoscopy", IEEE Wireless Communications and Networking Conference (WCNC), **(2015)**.

[4] G. Liu, Y. Geng and K. Pahlavan, "Effects of calibration RFID tags on performance of inertial navigation in indoor environment", 2015 International Conference on Computing, Networking and Communications (ICNC), **(2015)**.

[5] J. He, Y. Geng, Y. Wan, S. Li and K. Pahlavan, "A cyber physical test-bed for virtualization of RF access environment for body sensor network", IEEE Sensor Journal, vol. 13, no. 10, **(2013)**, pp. 3826-3836.

[6] W. Huang and Y. Geng, "Identification Method of Attack Path Based on Immune Intrusion Detection", Journal of Networks, vol. 9, no. 4, **(2014)**, pp. 964-971.

[7] G. Bao, L. Mi, Y. Geng, M. Zhou and K. Pahlavan, "A video-based speed estimation technique for localizing the wireless capsule endoscope inside gastrointestinal tract", 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), **(2014)**.

[8] D. Zeng and Y. Geng, "Content distribution mechanism in mobile P2P network", Journal of Networks, vol. 9, no. 5, **(2014)**, pp. 1229-1236.

[9] M. Zhou, G. Bao, Y. Geng, B. Alkandari and X. Li, "Polyp detection and radius measurement in small intestine using video capsule endoscopy", 2014 7th International Conference on Biomedical Engineering and Informatics (BMEI), **(2014)**.

[10] W. Gu, Z. Lv and M. Hao, "Change detection method for remote sensing images based on an improved Markov random field", Multimedia Tools and Applications, **(2015)**, pp. 1-16.

[11] Z. Lv, A. Tek and F. D. Silva, "Game on, science-how video game technology may help biologists tackle visualization challenges", PloS one, vol. 8, no. 3, **(2013)**, pp. 57990.

[12] Z. Chen, W. Huang and Z. Lv, "Towards a face recognition method based on uncorrelated discriminant sparse preserving projection", Multimedia Tools and Applications, **(2015)**, pp. 1-15.

[13] D. Jiang, X. Ying and Y. Han, "Collaborative multi-hop routing in cognitive wireless networks", Wireless Personal Communications, **(2015)**, pp. 1-23.

[14] Z. Lv, A. Tek and F. D. Silva, "Game on, science-how video game technology may help biologists tackle visualization challenges", PloS one, vol. 8, no. 3, **(2013)**, pp. 57990.

[15] D. Jiang, Z. Xu and Z. Lv, "A multicast delivery approach with minimum energy consumption for wireless multi-hop networks", Telecommunication Systems, **(2015)**, pp. 1-12.

[16] C. Fu, P. Zhang and J. Jiang, "A Bayesian approach for sleep and wake classification based on dynamic time warping method", Multimedia Tools and Applications, **(2015)**, pp. 1-20.

[17] Z. Lv, "Wearable smartphone: Wearable hybrid framework for hand and foot gesture interaction on smartphone", Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. IEEE, **(2013)**, pp. 436-443.

[18] Y. Lin, J. Yang and Z. Lv, "A Self-Assessment Stereo Capture Model Applicable to the Internet of Things", Sensors, vol. 15, no. 8, **(2015)**, pp. 20925-20944.

[19] J. Yang, S. He and Y. Lin, "Multimedia cloud transmission and storage system based on internet of things", Multimedia Tools and Applications, **(2015)**, pp. 1-16.

[20] Z. Lv, T. Yin and Y. Han, "WebVR - web virtual reality engine based on P2P network", Journal of Networks, vol. 6, no. 7, **(2011)**, pp. 990-998.

[21] J. Yang, S. He and Y. Lin, "Multimedia cloud transmission and storage system based on internet of things", Multimedia Tools and Applications, **(2015)**.

[22] C. Guo, Z. Liu and M. Jin, "The research on optimization of auto supply chain network robust model under macroeconomic fluctuations", Chaos, Solutions & Fractals, **(2015)**.

[23] X. Li, Z. Lv and J. Hu, "XEarth: A 3D GIS Platform for managing massive city information", Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2015 IEEE International Conference on. IEEE, **(2015)**, pp. 1-6.

Authors

Hu Xiaoyong, lecturer, master, graduated from Jishou University in 2011, the National Traditional Sports Science College. His main research direction is the national folk sports. Now He is a teacher in School of physical education Guiyang University.