THE APPLICATION OF INFORMATION ENTROPY THEORY BASED DATA CLASSIFICATION ALGORITHM IN THE SELECTION OF TALENTS IN HOTELS

A. Youyu Hu
Department of Mathematics
Qingdao Vocational and Technical College of Hotel Management
Qingdao China
hyy6001@163.com

Abstract. Background: With the rapid development of the society, some excellent enterprises are having a growing demand for talents. The recruitment and selection of talents is always concerned by the managers of enterprises. Materials and methods: This study analyzed the human resource of hotels and introduced C4.5 algorithm in decision tree algorithm and its implementation procedures. Talents were selected by investigating data of relevant ability of them and performing data analysis and data mining using C4.5 algorithm. Objective: Information entropy based data classi

cation C4.5 algorithm was used to screen hotel talents. Results: Decision tree was obtained by calculating the comprehensive ability, working experience, random response capability and psychological quality of the enrolled employees. During the talents selection in hotels, the C4.5 algorithm can make the process simple and convenient.Decision tree algorithm can be used in the selection of talents in enterprises to preprocess data, construct data mining model of talent selection, and solve problems appearing in the recruitment and selection of talents.

Keywords: Data mining, decision tree, information entropy, human resource, C4.5 algorithm, information entropy theory, competition, data analysis, talents selection.

Significance statement

This study investigated the application of C4.5 in the talents selection in hotels. This study investigated decision tree classification algorithm, aiming to promote the application of data mining technology in the hotel industry. The paper has no conflicts of interest.

1. Introduction

In recent years, the market competition in hotel industry has showed a tendency of information high-technicalization and globalization. Among all resources, human resource has become one of the most important competition factors. Hotel talents investigation and selection using network algorithms is more competitive. Decision tree algorithm has been extensively applied in the fields such as machine learning, knowledge discovery and human resource analysis [1]. Data Mining (DM) as a kind of data analysis method and technique of finding potential information among massive data has become a hot spot in all circles [2,

3]. Chen J. [4] discussed the shortcomings of ID3 decision tree and proposed an improved algorithm integrating ID3 and association function which effectively made up the defects of ID3 decision tree and obtained more reasonable and effective rules. Wang Y. H. [5] applied Gini coefficient decision tree algorithm in human resource management system, constructed the decision trees by taking working tasks, working quality and working attitude as the decision attributes, established the enterprise performance evaluation model, and verified its feasibility. Bai W. J. [6] et al. investigated the recruitment and employment of university students, established human resource information database for screening professional talents through k-mean algorithm, and made classification and prediction on the employment data of previous graduates using C 4.5 algorithm.

In decision tree algorithm, the complexity and classification accuracy of decision tree are the two most important factors for consideration. The commonly used evaluation indexes include prediction accuracy, i.e., describing classification models and predicting new or unknown data class accurately, description conciseness (the more concise the description of model is, the easier to understand), computational complexity, and the robustness of model and processing scale [7]. This study analyzed data mining for the investigation of hotel talents by using C4.5 algorithm on the basis of information entropy theory and proposed to rapidly select and reasonably assign excellent talents using informatization.

2. Analysis of hotel human resource

Human resource as the most important core competition factor in hotel development and management has significant influence on the competition between hotels. The key for hotels with service nature to become the top-ranking star hotels lies on the working enthusiasm of staffs [8, 9]. The intensive competition between hotels, essentially, is the competition of talents. Rich human resource can greatly promote the development of hotels; however, many problems remain to be solved in the selection and management of talents, for example, how to attract and select servant talents who are needed by hotels, systematically screen talents, and arrange optimal positions for the existing staffs to realize the values of people and create more benefits for hotels. During the allocation of human resource, the screening and allocation of talents based on the reasonable use of information technology algorithm includes two parts [10]. The first part is the initial allocation of human resource, i.e., making a rough determination on the interviewees according to the provided data and rough impression in interview. The second part is the reallocation of human resource, i.e., investigating and selecting excellent talents using informatization algorithm and arranging staffs to suitable positions after assessment.

3. C4.5 Decision Tree Algorithm

3.1 The overview of decision tree

Decision tree is a process of classifying data through a series of rules, which is a tree structure that can classify data automatically and is the knowledge

representation of tree structure; it can be converted into decision rules directly [8-10]. Decision tree learning means comparing attribute values in the internal node adopting the top-down recursive method, determining the branches below the node according to different attribute values, and coming to conclusions in leaf nodes of a decision tree. The construction process of decision tree is shown in figure 1. As to data classification, decision tree is useful as it can construct a tree for modeling for classification problem and the tree can effectively classify problems and reach classification results. Therefore, two steps are needed in the classification using decision tree, i.e., establishing a classification model and applying data sets into the model for classification.

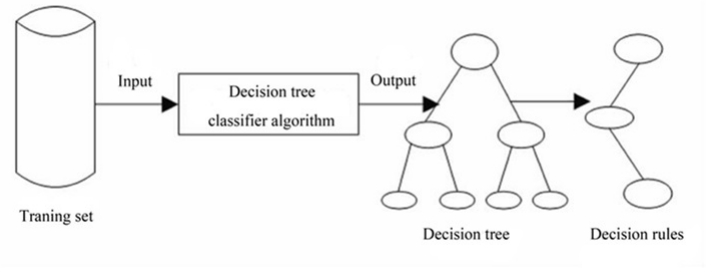

Figure 1: The construction process of decision tree

3.2 C4.5 Algorithm

The construction process of decision tree is a process of segmenting data uninterruptedly, and each segment is corresponding to a node. The C4.5 algorithm forms a decision tree recursively based on information entropy method.

(1) Definition of entropy measurement

Entropy is a metric that is widely used in information theory. Entropy with respect to Boolean classification is defined as:

$$Entropy(S) \equiv -P\log 2P - P\log 2P[1] \tag{3.1}$$

where P stands for the ratio of positive example in S and $-P$ stands for the ratio of counterexample in S. If the target attribute has c different values, then the entropy is defined as:

$$Entropy(S) \equiv \sum_{i=1}^{c} -Pi\log 2Pi \tag{3.2}$$

(2) Definition of information gain

Information gain metrics of training data of attributive classification can be defined by giving a definition for entropy:

$$Gain(S,A) \equiv Entropy(S) - \sum_{v\in Values(A)} \frac{|S_V|}{S} Entropy(S_V) \tag{3.3}$$

Values (A) are all the possible subsets of attribute A; S_V is the subset of V value of attribute A in S. (3) Definition of split information and the ratio of gain Split information:

$$SplitInformation(S,A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log 2 \frac{|S_i|}{|S|} \tag{3.4}$$

Gain ratio:

$$GainRatio(S,A) \equiv \frac{Gain(S,A)}{SplitInformation(S,A)}. \tag{3.5}$$

The specific steps of the C4.5 algorithm are as follows: Step 1: Data source are preprocessed, and a training set of decision tree is formed by processing discretization for continuous attribute variables (ignore it if there is no continuous valued attributes) [12]. Firstly, the minimum value a_0 and maximum value $a_n + 1$ of the continuous attribute are searched on the basis of the original data; secondly, n values are inserted into the interval [a; b] to divide it into $n + 1$ sections; thirdly, a_i; $i = 1, 2, \ldots, n$ is taken as the segment point to divided interval $[a_0, a_{n+1}]$ into two subsections: $[a_0, a_i]$ and $[a_{i+1}, a_{n+1}]$ which are corresponding to two kinds of values of the continuous attribute variable. There are n division modes.

Step 2: The information gain and ratio of information gain ratio of each attribute is calculated. Firstly, the information gain Gain (A) of attribute A is calculated. Secondly, the ratio of information gain Gain Ratio (A) of attribute A is calculated.

The attribute with the largest information gain ratio was selected as the current attribute node to get the root node of decision tree.

Step 3: Each value of root node is corresponding to a subset, and the second step process is performed recursively for sample subset until the data has the same value in classification attributes in each subset. Finally, a decision tree is generated [13].

Step 4: New data sets are classified according to the extraction and classification rules of the constructed decision tree.

4. Application of decision tree in the selection of talents

With the continuous development of computer technology, C4.5 algorithm has been successfully applied in the business, financial, health care and other fields [11, 12]. Currently many universities and organizations are facing with the problem of talents selection. Colleges and universities will select some excellent talents to attend various national or provincial competitions; organizations tend to choose comprehensive talents. Whether a student is suitable for an organization or whether he can make achievements in a college or university needs to be determined. In order to solve such problems, this study introduced the data

classification algorithm - C4.5 algorithm [13] in the selection of talents based on information entropy theory by taking the selection of excellent talents in hostels as an example.

First of all, all staffs who achieved ranking in the selection of hotel talents were prepared (table 1), and the attributes that are suitable for digging were determined.

Name	Average score of comprehensive ability	Work experience	Reaction capability	Psychological quality	Whtether ranking or not achieved
Li Yi	A	B	A	C	Yes
An Ji	C	A	B	A	Yes
Mao Zhenyu	B	B	C	B	No
...	...	...	...	...	...

Table 1: Relevant data of the staffs who achieved ranking

Value ranges of data attributes were as follows: Attribute (Average score of comprehensive ability) = $\{A, B, C\}$; attribute (Work experience) = $\{A, B, C\}$; attribute (Reaction capability) = $\{A, B, C\}$; attribute (Psychological quality in game) = $\{A, B, C\}$;

$$Entropy(S) \equiv -P\log 2P - P\log 2P \equiv 1.226 \tag{4.1}$$

Information gain, split information and ratio of information gain of each attribute were obtained through calculation:

Gain(S, Average score of comprehensive ability) = 0.5099; gain (S, Work experience) 0.3061; gain (S, reaction capability) = 0.8835; gain (S, psychological quality in game) $\equiv$ 0.5594.

Gain Ratio (S, average score of computer major) $\equiv$ 0.236331; gain Ratio (S, experience in game) $\equiv$ 0.021109; gain Ratio (S, reaction capability) $\equiv$ 0.170 241; gain Ratio (S, psychological quality in game) $\equiv$ 0.105557.

According to information gain of four kinds of attributes, the ratio of information gain of average score attribute of comprehensive ability is the maximum, therefore; average score attribute of comprehensive ability was selected as the root node of decision tree. Then the following attributes were calculated: Gain Ratio (grade A, working experience) $\equiv$ 0.102021; Gain Ratio (grade A, reaction capability) $\equiv$ 0.104401; Gain Ratio (grade A, psychological quality) $\equiv$ 0.349101; Gain Ratio (grade B, work experience) 0.220584; Gain Ratio (grade B, reaction capability) $\equiv$ 0.404112; Gain Ratio (grade B, psychological quality) $\equiv$ 0.01201; ... ... therefore, psychological quality was regarded as the node under grade A and reaction capability as the node under grade B. Finally a decision tree was obtained.

Potential talents selection rules were digged out from the vast and clutter data using C4.5 algorithm based on demands, which omits the cumbersome

steps of human analysis and let the system analyze a large amount of data in the database automatically to obtain information [14, 15]. Due to the various existing data categories and large data size, human is very difficult to find useful information from those data. However, C4.5 algorithm that can perform dispersing processing on continuous data and reach highly-accurate results is suitable for analyzing that kind of data. Excellent servant talents could be efficiently screened through C4.5 algorithm to make managers understand the condition of human resource better. decision tree.

5. Discussion

With the rapid development of Chinese economy, the competition between hotels becomes more and more intense. Human resource as the core competition factor of hotels plays an important role in hotel competition. In talents recruitment and selection, the reasonable application of data informatization technology makes the process more convenient. Bouajaja S. et al. [16] proposed solving the problem of human resource optimal distribution using ant colony optimization which could improve the productivity and market competition of enterprises. Mukherjee A. N. et al. [17] studied applicant tracking system and proposed human resource information system model by applying applicant tracking system in human resource management. Based on the decision tree algorithm in data classification algorithm, this study analyzed relevant data of hotel staffs. Compared to other algorithms, C4.5 algorithm can process big data with continuous data discretization.

Evaluating the performance of the staffs with C4.5 algorithm could find out attributes that were suitable for mining and accurately deduce their values to the hotel. Then to achieve more benefits, the staffs were allocated and a correct human resource short-term plan was formulated according to the deduced information. There were some defects in this study. For example, the study did not involve all aspects, the classification of decision attributes was not comprehensive, and there was no comparison with other algorithms, which are expected to be improved in the future.

6. Conclusions

In conclusion, the informatization of human resource management can make the selection of talents more efficient and improve the competition of hotels. C4.5 algorithm, a data classification algorithm based on information entropy theory, was used to calculate different decision attributes in the process of talents selection. The algorithm makes talents selection convenient and creates beneficial prerequisite conditions for the success of hotels in market competition.

References

[1] J. Chen, D.L. Luo, F.X. Mu, *An improved ID3 decision tree algorithm*, Advanced Materials Research & Data Analysis, 2014, 962-965:2842-2847.

[2] R.J. Simpson, S. Sidhar, T.J. Peters, *A critical survey of data grid replication strategies based on data mining techniques*, Procedia Computer Science, 51 (1) (2015), 2779-2788.

[3] I.H. Witten, E. Frank, *Data mining: practical machine learning tools and techniques*, Biomedical Engineering Online, 51 (1) (2011), 95-97.

[4] J. Chen, D.L. Luo, F.X. Mu, *An improved ID3 decision tree algorithm*, Advanced Materials Research, 2014, 962-965:2842-2847.

[5] Y.H. Wang, K.B. Wang, H.Y. Xue, W.H. Zhou, *Application of Gini Index decision tree algorithm to the performance management of human resources*, Journal of Yunnan University of Nationalities, 2013.

[6] W. J. Bai, *Cultivation Model of Human Resource Management Based on Internet*, Advanced Materials Research, 2014, 971-973: 2305-2308.

[7] M. S. Yorgun, R.B. Rood, *A decision tree algorithm for investigation of model biases related to dynamical cores and physical parameterizations,* Journal of Advances in Modeling Earth Systems, 2016.

[8] Y.S. Chauhan, N.K. Patel, *Human Resources Management practices and job satisfaction: a study of hotel industry*, Abhinav-National Monthly Refereed Journal of Research in Commerce & Management, 2014.

[9] K. Orfin, M. Sidorkiewicz, A. Tokarzkocik, *Human resource management in chain hotels on the example of Radisson Blu hotel in Szczecin*, Scientific Journal, 31 (2015), 287-302.

[10] L. He, J. Qi, *Enterprise Human Resources Information Mining Based on Improved Apriori Algorithm*, Journal of Networks, 8 (5) (2013), 1138.

[11] Z.N. Chang, *The application of C4.5 algorithm based on SMOTE in financial distress prediction model*, Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on. IEEE, 2011, 5852-5855.

[12] J. Zhang, S. Wang, *Study on Application of C4.5 Algorithm in Medical Insurance Data Mining*, Journal of Shijiazhuang Railway Institute, 2008, 5852-5855.

[13] I.D. Nurpratami, I.S. Sitanggang, *Classification Rules for Hotspot Occurrences Using Spatial Entropy-based Decision Tree Algorithm*, Procedia Environmental Sciences, 24 (2015), 120-126.

[14] H. Jantan, A.R. Hamdan, Z.A. Othman, *Talent knowledge acquisition using data mining classification techniques*, Data Mining and Optimization. IEEE, (2011), 32-37.

[15] I.A. Kareem, M.G. Duaimi, *Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization*, International Journal of Computer Science & Mobile Computing, 3 (6) (2014), 176-183.

[16] S. Bouajaja, N. Dridi, *Research on the optimal parameters of ACO algorithm for a human resource allocation problem*, Ervice Operations and Logistics, and Informatics, 2015.

[17] A.N. Mukherjee, S. Bhattacharyya, B. Risika, *Role of Information Technology in Human Resource Management of SME: A Study on the Use of Applicant Tracking System*, Ibmrds Journal of Management & Research, 2014.

Accepted: 1.03.2017