Journal of Contemporary Educational Research, 2023, Volume 7, Issue 2
http://ojs.bbwpublisher.com/index.php/JCER
ISSN Online: 2208-8474
ISSN Print: 2208-8466

Research on the Intelligent Distribution System of College Dormitory Based on the Decision Tree Classification Algorithm

Huiping Han[1], Beida Wang[2]*

[1]China University of Geosciences (Beijing), Beijing 100083, China

[2]School of Earth Sciences and Engineering, Nanjing University, Nanjing 210023, Jiangsu Province, China

***Corresponding author:** Beida Wang, 951551918@qq.com

Abstract: The trend toward designing an intelligent distribution system based on students' individual differences and individual needs has taken precedence in view of the traditional dormitory distribution system, which neglects the students' personality traits, causes dormitory disputes, and affects the students' quality of life and academic quality. This paper collects freshmen's data according to college students' personal preferences, conducts a classification comparison, uses the decision tree classification algorithm based on the information gain principle as the core algorithm of dormitory allocation, determines the description rules of students' personal preferences and decision tree classification preferences, completes the conceptual design of the database of entity relations and data dictionaries, meets students' personality classification requirements for the dormitory, and lays the foundation for the intelligent dormitory allocation system.

Keywords: Intelligent allocation; Personal preference; Information gain; Decision tree classification; Individualization

Online publication: February 1, 2023

1. Introduction

In 2019, the scale of enrollment in ordinary colleges and universities reached 112.6 million, and the number of teaching staff reached 2.5667 million. Facing such a large service body, logistics management is faced with several major challenges. The traditional dormitory distribution system based on majors and reporting order deviates from the policy guidance of "building a high-quality education system." Giving full consideration to students' work and rest habits, health habits, personality traits, and other attributes, a study is carried out on the dormitory intelligent distribution system.

2. Design and analysis of the intelligent distribution system

2.1. Demand analysis

Pre-survey – set questionnaire content – publish questionnaire – collect preference data – import data to ensure the accuracy, effectiveness, and pertinence of the intelligent distribution system. Complete the initial allocation – receive feedback – optimize allocation – form a smart feedback system. Through the information platform, real-time viewing, statistics, addition, deletion, modification, and other operations can be done.

In order to meet the management requirements of the school and the individual development of students, the following principles should be adhered to when allocating dormitories: students of the same grade,

major, and class shall be allocated to the same dormitory and floor as much as possible; and students with high preference matching shall be assigned to the same dormitory.

2.2. Module design

The function module is divided into three parts. The first part is the preference information input section. Input the preference information into the database and preprocess the missing and duplicate values. The second part is the intelligent distribution section. After receiving the preference data, based on the decision tree classification algorithm, select the allocation scheme with the highest overall harmony and realize the personalization and automation of dormitory allocation. The final part is the distribution result intelligent feedback module. Allow students to exchange information with the system administrator according to the allocation results, form a satisfaction feedback report, as well as improve and optimize the functional areas.

2.3. Process analysis

Complete the questionnaire's question setting according to the hard and soft constraints. Collect data – input – intelligent allocation – publicity results – feedback – improvement – generate a satisfaction feedback form.

3. Preference-based allocation model

3.1. Questionnaire settings

In the new era, college students who prefer online interaction comprise the majority of students on campus. Meeting their diversified needs and properly handling the conflicts caused by living habits have become important issues of student management. As early as 2017, China Youth Network launched an investigation on the dormitory relationship of 958 college students nationwide. The results showed that more than 40% of students had disputes and conflicts with their roommates, and nearly 30% of them had thought of changing their dormitories, of which living habits were the main factor. At present, Shanghai Jiaotong University, Shanghai University, and Nankai University have been experimenting with the "freshman dormitory self-selection" distribution system.

Therefore, the design of a targeted, accurate, and comprehensive questionnaire that can distinguish and meet the individual needs of college students is an important premise of the "freshman dormitory self-selection" distribution system. The impact of negative dormitory incidents on students can be divided into eight aspects: comfortability, emotional belonging, general sharing, personal safety, financial security, respect, public affairs, and interpersonal independence. Conventional coping methods include neglect, patience, ridicule, understanding and tolerance, arguments, criticism, self-reflection, communication, and reminder. According to individual needs, there are 10 factors that affect dormitory relationships: the gap between the rich and poor, different provinces, campus experience, work and rest time, snoring in sleep, bad odor, personal hygiene, cards and games, preference for sports, as well as internal and external traits. Therefore, the reasons for college students' dissatisfaction with their accommodation can be summarized as follows: difference in living habits, difference in personality, difference in interests, and difference in economic situation. According to these four differences, the acceptance level of the students in the questionnaire is set as follows: mind very much, somewhat mind, neutral, do not really mind, do not mind.

3.2. Algorithm analysis

The dormitory allocation model can be divided into unsupervised classification and supervised classification. The relationship between data can be obtained according to clustering or certain models via the unsupervised classification. The algorithm has no clear target, and the data in the data set do not need to be labeled. Hence, the effect is difficult to evaluate. On the other hand, the correlation between input and

output results can be obtained according to the existing data set via the supervised classification, and an optimal model can also be obtained through training. The supervised classification judges the optimal label for the data with only features without label and has a clear classification objective. The classification effect is easy to evaluate. In dormitory allocation, the classification results need to be evaluated, and the training model needs to be supervised, so the supervised classification is selected. In the supervised classification, the decision tree algorithm is easy to understand. It can be used for small data sets and is suitable for the classification of discrete data. It also conforms to the characteristics of student preference data set. Decision tree is a basic classification and regression method. Starting from the root node of the whole decision tree, the sub-tree shall be judged for further decision-making and decomposition according to the characteristics of each node. New features are adopted in each decision-making process. Appropriate decision threshold can reduce the classification error rate to a low level until it points to leaf nodes or no longer needs to make decisions. In that way, the density of data value can be improved as much as possible in a relatively large amount of data [1].

The decision tree generation process is transparent and visible. There is no need to preprocess the data too much, and the workload is low. It is also not affected by missing values. The decision tree is divided into categorical variables and continuous variables. Since the preference data does not meet the requirements of continuous variables, categorical variables are selected. In order to better determine the feature selection principle, the feature selection method based on information gain is introduced. Information gain refers to the change of information before and after data set division. The information gain obtained by calculating each characteristic value of data set division is the best choice. Information entropy can be defined here. On the premise that things to be classified are in multiple classes, the information x_i can be defined as follows:

$$l(x_i) = -log_2 p\,(x_i),$$

where $p(x_i)$ is the probability of selecting the classification. In order to calculate the entropy, the expected value of information contained in all values for all classes needs to be calculated:

$$H = -\sum_{i=1}^{n} p(x_i)\, log_2 p\,(x_i),$$

where *n* is the class number, and the value of entropy is positively correlated with the uncertainty of random variables. When the probability of entropy is obtained by data estimation, the corresponding entropy is called empirical entropy. Setting the training data set *D*, the empirical entropy of the training data set *D* will be as follows:

$$H(D) = -\sum \frac{|c_k|}{|D|} log_2 \frac{|c_k|}{|D|}.$$

Let there be *n* categories (k = 1, 2, ... n), with $|c_k|$ representing the c_k sample size belonging to the class, and $|D|$ representing the sample size of training data set *D*.

Before introducing information gain, the concept of conditional entropy must be defined. Information gain can represent the degree to which the information uncertainty of class *Y* is reduced when the information of feature *X* is known; conditional entropy $H(Y|X)$ represents the uncertainty of random variable *Y* under the condition that random variable *X* is known. The mathematical expectation of information entropy for *X* of conditional probability distribution of random variable *Y* under given conditions is defined as follows:

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i),\ p_i = P(X = x_i).$$

If the probability is zero, $0\,log\,0 = 0$. In summary, information gain can be defined as follows: $g(D,C) = H(D) - H(D|C)$. The information gain $g(D,C)$ of feature *C* on training data set *D* is defined as the difference between the empirical entropy $H(D)$ of training data set *D* and the empirical conditional entropy $H(D|C)$ of *D* under the given condition of feature *C*. However, the size of the information gain value is closely related to the relative size of training data set D, which is a relative value. When the classification of the decision tree is difficult, that is, when the empirical entropy is large, the information gain value is larger, and vice versa. In order to eliminate the relative relation between information gain and empirical entropy, gain experience ratio $g_R(D,C)$ is introduced: $g_R(D,C) = \frac{g(D,C)}{H(D)}$.

3.3. Model building

According to the decision tree classification and information gain method, define the preference state set *P*, decision attribute set *D*, decision attribute value set *V*, decision attribute ordering set *S*, and preference state description rule *f* of the subject to be assigned. According to the preferences of the students to be assigned, "mind the difference in living habits very much," "do not mind the difference in economic situation," *etc*., let the number of preference states be *n*, $P = \{s_1, s_2, s_3 \ldots, s_n\}$. Among the attributes of "mind personality differences very much" are "prefer more active roommates rather than introverted roommates" and "prefer more talkative roommates." Let the number of decision attributes be *m*, $D = \{d_1, d_2, d_3 \ldots, d_m\}$. According to the decision attribute value of "I hope my roommate does not smoke," it can be divided into acceptance and rejection. Assume that there are *k* attribute values corresponding to the decision attribute; $V = \{v_1, v_2, v_3 \ldots, v_k\}$. Decision attribute ranking set S has the same number of elements as the decision attribute set D, $S = \{s_1, s_2, s_3 \ldots, s_m\}$. s_x is the position of the decision attribute whose order is *x*($x \in [1, m]$) in decision attribute set D. Therefore, the sorted decision attribute set is: $D' = \{d_{s_1}, d_{s_2}, d_{s_3} \ldots, d_{s_m}\}$. Here, the expert analysis method and the decision attribute ranking method are integrated as the ranking method of each decision attribute, that is, the decision tree classification algorithm is used as the auxiliary basis of decision attribute ranking. Preference state description rule *f* represents the relationship between the preference state of the subject to be assigned and the value of each decision attribute, that is, $P(t) = f(V_{s_1}(t), V_{s_1}(t), \ldots, V_{s_m}(t))$; among them, $P(t) \in P, Z_{a_i}(t) \in Z_{a_i}, i \in [1, m]$.

3.4. Process settings

During model analysis, the preference state of the subject to be assigned and the value and ranking of decision attributes that influence the subject to be assigned to reach a certain preference state are clarified; the classification decision tree is finally generated [2].

Set the model to form a set by analyzing the preference. The preference state set *P* is obtained; the decision attribute set and its corresponding value category are determined; and the decision attribute set *D* and decision attribute value set *V* are obtained.

Sort the allocation decision attributes by analyzing the information gain and the Gini impurity of decision tree classification. A more appropriate decision tree classification and more accurate decision attribute ranking results are obtained, and the decision attribute ranking set S under different decision tree depths is determined.

In the setting and visualization of the decision tree, according to the score of S and both model training and test sets, the optimal depth is selected for training, and the intelligent allocation decision tree model is obtained.

When forming preference state description rules, the formation rules of preference state are summarized according to the generated decision tree.

3.5. Model analysis

The survey data was preprocessed before model analysis, and the classification of four qualitative variables in the data set was encoded. However, numerical variables could only be identified by the model as size relations, so dummy variables were introduced to replace different categories of qualitative variables with one or more qualitative variables with positive integers (such as 1, 2, 3, and 4) [3]. For qualitative variables such as life habits, personality traits, interests, and economic conditions, set the acceptance level as follows: 1, mind very much; 2, somewhat mind; 3, neutral; 4, do not really mind; 5, do not mind.

3.5.1. Set the model component set

The survey questions are set as living habits, personality traits, interests, and economic conditions, and the preference state set *P* of the subject to be assigned can be defined as follows:

$$P = \{Habits, personality\ traits, interests, finances\}.$$

Define four decision attributes V according to the pre-survey results.

$$V_{Acceptance\ of\ lifestyle\ differences} = \{1,2,3,4,5\}$$

$$V_{Acceptance\ of\ personality\ differences} = \{1,2,3,4,5\}$$

$$V_{Acceptance\ of\ interest\ differences} = \{1,2,3,4,5\}$$

$$V_{Acceptance\ of\ economic\ differences} = \{1,2,3,4,5\}$$

3.5.2. Sort allocation decision attributes

Decision attribute sorting based on decision tree classification is the main method, supplemented by expert analysis method. Two separation principles, information gain and Gini impurity, were selected in the decision tree classification algorithm, respectively. The depth of the decision tree was taken as the independent variable, and the scores of the training set and the test set were taken as the dependent variable. The differences in scores of the two data sets were observed, as shown in **Table 1**.

Table 1. Data set score variance

Data set score variance					
Data set scores based on Gini impurity			**Data set score based on information gain**		
max_depth	**train_score**	**test_socre**	**max_depth**	**train_score**	**test_socre**
2	0.67619	0.688889	2	0.67619	0.688889
3	0.67619	0.688889	3	0.67619	0.688889
4	0.704762	0.711111	4	0.714286	0.666667
5	0.752381	0.733333	5	0.752381	0.711111
6	0.809524	0.755556	6	0.828571	0.755556
7	0.866667	0.8	7	0.885714	0.822222
8	0.895238	0.911111	8	0.895238	0.911111
9	0.942857	0.955556	9	0.942857	0.933333
10	0.980952	0.933333	10	0.971429	0.933333
11	0.990476	0.955556	11	0.980952	0.955556

The differences in scores of the data sets are visualized in **Figure 1**. In the decision tree classification simulation, the score of the training set and the test set is positively correlated with the maximum depth. When the maximum depth of the decision tree is 10 and 11, the score of the training set and the test set is above 0.97. Since the depth of the decision tree is large at this time, overfitting easily occurs, so the score of the test set is taken as the main selection basis. In order to prevent overfitting of the decision tree, the maximum depth of the decision tree classification model is set at 11. Following the algorithm principle comparison of student preference data, information gain, as the classification principle, has better fitting effect and better dormitory intelligent allocation effect [4], and the importance ranking of decision attributes is obtained: personality traits 0.30027634, life habits 0.29015753, economic situation 0.24953269, hobbies 0.16003344.

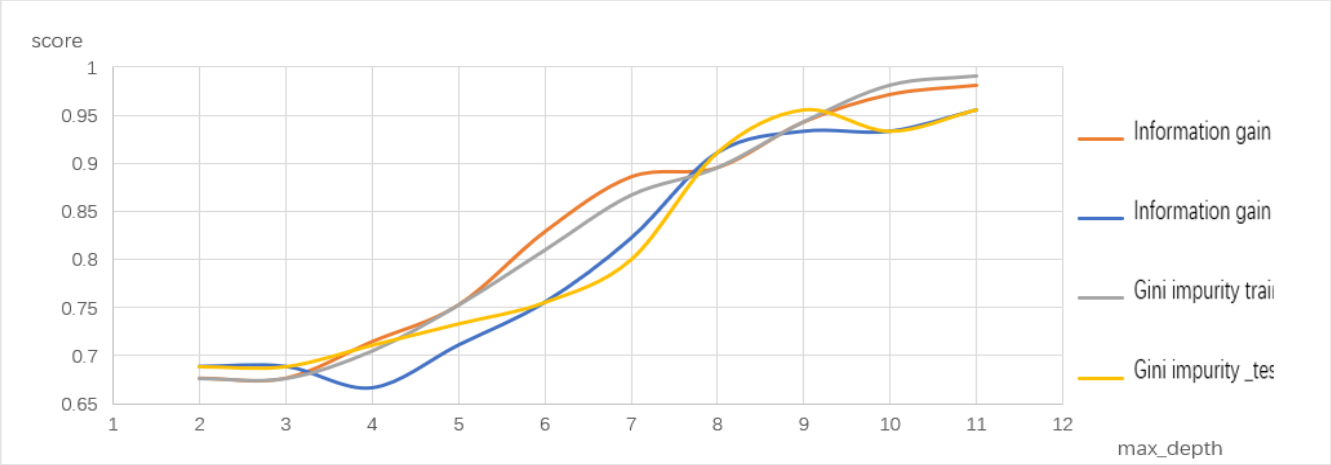

Figure 1. Data set scores of different classification principles

3.5.3. Setting and visualization of decision tree

According to the preference state set *P*, decision attribute set *D*, decision attribute value set *V*, and decision attribute ranking set *S*, parameters of the decision tree algorithm are set. In the classification results, the sample size of this layer of leaf nodes is small, and the entropy reaches a low level, indicating that most samples have completed the classification task well under the premise of eliminating overfitting.

3.5.4. Form preference state description rules

According to the nature of the decision tree classification algorithm, the path from the root node to any leaf node is the classification process of a certain sample, including the classification path based on the relationship between the size of each parameter value. Then, extract the preference state description rule *f*. Among them, Hobby refers to interest difference acceptability, Habit refers to living habit difference acceptability, Character refers to personality trait difference acceptability, and Consumption refers to economic situation difference acceptability.

```
|--- Hobby <= 3.50
|    |--- Habit <= 1.50
|    |    |--- Character <= 1.50
|    |    |    |--- Consumption <= 2.50
|    |    |    |    |--- class: 1
|    |    |    |--- Consumption > 2.50
|    |    |    |    |--- class: 2
```

```
|   |   |--- Character > 1.50
|   |   |   |--- Consumption <= 1.50
|   |   |   |   |--- class: 2
|   |   |   |--- Consumption > 1.50
|   |   |   |   |--- Consumption <= 3.50
|   |   |   |   |   |--- Habit <= 0.50
|   |   |   |   |   |   |--- class: 2
|   |   |   |   |   |--- Habit > 0.50
|   |   |   |   |   |   |--- Character <= 4.50
|   |   |   |   |   |   |   |--- Hobby <= 0.50
|   |   |   |   |   |   |   |   |--- class: 3
```

4. Database design

The main body of the intelligent distribution system is set as students, users, questionnaires, dormitory buildings, dormitory rooms, and system administrators. Along the "user – intelligent distribution system – system administrator" line description, a closed-loop operation is carried out.

Table 2 shows the data dictionary based on entity relationship.

Table 2. Category information

Serial number	Field name	Type	Length	Nonempty constraint	Description
Student information	id	varchar	10	Not null	Primary key_student number
	name	varchar	10	Not null	student's name
	state	varchar	2	Not null	whether to fill in the questionnaire
Dormitory information	id	varchar	10	Not null	Primary key_dorm number
	stu_num	int	/	/	capacity
Dormitory building information	id	varchar	10	Not null	Primary key_building number
	level_num	int	/	Not null	number of floors
User information	id	varchar	20	Not null	Primary key_username
	password	varchar	20	Not null	Password
System administrator information	super_id	varchar	20	Not null	Primary key_username
	super_password	varchar	20	Not null	password
College information	id	varchar	10	Not null	Primary key_college number
	name	varchar	10	Not null	college name
Student preference investigation questionnaire information	id	Int	10	Not null	Primary key_problem number
	habit	varchar	2	Not null	living habit
	character	varchar	2	Not null	personality traits
	hobby	varchar	2	Not null	hobbies and interests
	consumption	varchar	2	Not null	economic situation

5. Conclusion

The dormitory intelligent distribution system can reduce dormitory management energy consumption. It is one of the important issues in green campus construction. This paper is devoted to the study of a dormitory allocation system preferred by students. Using the decision tree classification algorithm as the core algorithm of intelligent allocation is an exploration and innovation of the traditional dormitory allocation

system function, but also a bold attempt to people-oriented dormitory management function, in order to contribute to the field of college dormitory allocation research.

Disclosure statement

The author declares no conflict of interest.

References

[1] Dong N, Mou Y, Liu Y, 2019, Digital Campus Construction Based on Big Data Era. Electronic Technology and Software Engineering, 21: 163–164.

[2] Yin X, Sun P, Han B, 2020, Behavior Rule Modeling of Combat Entity Based on Decision Tree. Command, Control and Simulation, 42(01): 15–19.

[3] Ge Z, Li Y, 2020, Dormitory Centralized Management Allocation Scheme Based on Integer Programming and 0-1 Backpack Problem: A Case Study of Guilin University of Electronic Science and Technology. Information and Computer (Theoretical Edition), 32(17): 3–5.

[4] Li M, Xiao L, 2020, Simulation Research on Intelligent Allocation of College Students Dormitory by Improved Genetic Algorithm. Computer Simulation, 37(07): 224–228.