Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 2, No. 2, 2024

Application of Drug Allocation Treatment Based on Decision Tree Algorithm

Weijie Xu, Jincheng Ge

School of Mathematical Sciences, Suzhou University of Science and Technology, Suzhou, Jiangsu, China

Abstract: Drug testing has always been an effective way to test drug effects and reduce drug costs. It is also a necessary measure for testing in the clinical stage. The results of drug testing will provide an important basis for the listing of new drugs, which is of great significance for promoting the development of medical science and improving the quality of life of patients. This article uses the decision tree algorithm to achieve effective allocation of drugs, improve the accuracy of treatment, determine and quantify the patient's key parameter information, including gender, blood pressure, and cholesterol. The model introduces two kinds of algorithm forms of decision tree through information entropy and Gini coefficient, and divides the data set into five kinds of decision trees through the algorithm, and the accuracy values all reach above 85%, and wait until the approximate accuracy effect under the condition of specifying random numbers, which provides scientific analysis for analyzing patients' choice of drugs.

Keywords: Decision tree; drug allocation; Gini coefficient.

1. Introduction

Development and testing of new drugs is a long, complex, and expensive process, and it is an important topic in the fields of bioinformatics and medicine. So far, from the beginning of research and development to the final approval and successful marketing of a new drug, its research and development costs have continued to rise in recent years, while the approval rate for marketing has only decreased but not increased [1]. According to documents released by the U.S. FDA in 2004, pharmaceutical product development costs have continued to increase since 2000, soaring 55% in 5 years, and future development costs will continue to expand. The current cost from research and development to marketing is approximately US$ 8 to 1.7 billion [2] . It can be seen that how to effectively utilize the acquired data samples and modern medical technology has become a core issue in research and development in improving drug treatment effects, reducing side effects, and reducing drug costs.

With the continuous development of science and technology, people have higher and higher requirements for the efficacy and safety of drugs. Therefore, new drugs need to be continuously developed to meet the challenges of diseases. Through the research and development of new drugs, more precise targeted drugs and more effective chemotherapy drugs can be developed to provide patients with better treatment options and improve the success rate of disease treatment. In the process of research and development, clinical testing of drugs is an essential part. Through clinical testing, the efficacy and safety of drugs on patients can be evaluated, possible side effects and risks can be discovered and resolved in a timely manner, and the quality and effectiveness of drugs can be guaranteed. Through clinical testing, the success rate of new drugs on the market can be improved, risks and costs for patients can be reduced, and new drugs can play their due role in clinical practice.

The main contributions of the article are as follows: First, it introduces the background meaning, basic concepts and analysis of the decision tree algorithm; then, it introduces the parameter optimization method of the decision tree and analyzes the characteristics of the decision tree algorithm; finally, it introduces application cases based on the decision tree algorithm, drug classification is used to predict classes for unknown patients.

2. Decision tree algorithm

2.1. Background

Decision tree is a popular machine learning algorithm used for classification and predictive modeling. It originated from the paper "Experiments in Induction" published by EB Hunt et al. in 1966 [3] . With the development of technology, decision trees have gradually become a popular algorithm for machine learning and are used in various aspects of research. The application scenarios of the decision tree model are very wide, covering all aspects, providing effective help for model establishment and optimization, and analysis of results. Luo Jia and Li Mingming [4] used the decision tree algorithm to predict the employment situation of college graduates, and provided targeted employment guidance programs through the results;

Yu Xiaolin and He Kaiping [5] studied the grouping case combination scheme related to disease diagnosis of cerebral hemorrhage patients through decision tree model, providing reference for relevant departments to formulate the grouping scheme and payment standard for localization of hospitalization costs of this disease.

Aiming at the demand analysis problem of travel insurance, Feng Yangyi [6] analyzed the forecast of tourists' demand in purchasing travel insurance through decision tree model. The application of decision tree in all walks of life just shows the strong applicability and wide application of decision tree model.

2.2. Algorithm description

The decision tree model is a classification and regression method based on a tree structure. It uses each node of the tree to represent features and divides the data by gradually selecting optimal features to achieve the purpose of hierarchical classification and regression. In the construction

process of the decision tree, a top-down greedy algorithm is generally used to recursively select the optimal features for segmentation until the stopping condition is met. In classification problems, the decision tree model can use the Gini index or information gain to evaluate the importance of features to perform segmentation; in regression problems, the decision tree model can use the mean square error to evaluate the importance of features to perform segmentation.

The advantages of the decision tree model are that it is intuitive and easy to understand, easy to explain, can handle classification and regression problems, and is insensitive to outliers. However, it also has the disadvantages of being easy to overfit and highly dependent on training data and feature selection instability. In practical applications, it is necessary to select an appropriate decision tree model based on specific problems and data characteristics, and improve the generalization ability and stability of the model through methods such as pruning and ensemble learning.

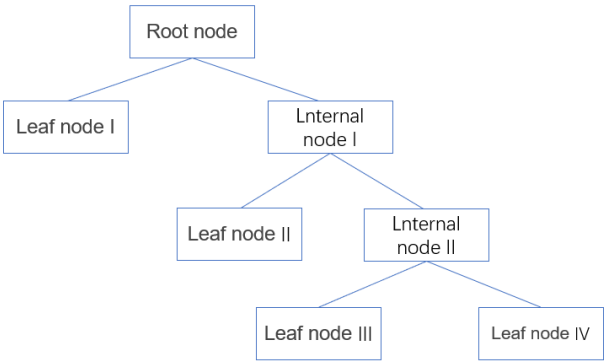


Figure 1 The basic idea of decision tree

2.3. Loss function of decision tree algorithm

Among many decision tree algorithms, the ID3 algorithm is the representative of the decision tree algorithm, and most of the algorithms are improved and implemented based on it. Decision tree models usually use Gini index or information gain as the loss function. The Gini index is a measure of the purity of a data set, while information gain is a measure of the information gain brought by features to the model. When building a decision tree model, the optimal features are selected to divide the data set by minimizing the Gini index or maximizing the information gain, thereby establishing a decision tree model.

2.3.1. Information gain

For the ID3 algorithm, information gain is used as the selection criterion when selecting attributes at all levels of the decision tree, so as to obtain the maximum category information of the tested record.

Suppose S is a set of s data samples, assume that the class label attribute has m different values, and define m different classes $C_i(i = 1,2,\ldots,n)$. Assuming to be $S_i$the number of samples in the class C_i, the expected information required to classify a given sample is:

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

Among them, $p_i = S_i/S$is the probability that any sample belongs toC_i .The logarithmic function has base 2 because the information is encoded in binary.

Suppose attribute *A* has *v* different values $\{ a_1, a_2, ..., a_v\}$. *S can* be divided into *v* subsets $\{ S_1, S_2, ..., S_v\}$ $S_j$by attribute *A, where* the samples in have the same value a (j =1,2,..., v) on attribute *A*.

Let S_i, $j$be the number of samples of the class in the subset S_jC_i. The entropy or information expectation of partitioning *A into subsets is:*

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

The smaller the entropy value, the higher the purity of the subset division. For a given subset S_j, its information expectation has the formula:

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

Among them, $s_{ij} = \frac{s_{ij}}{|s_j|}$is the probability that the sample S_j belongs to C_j, and the information gain obtained by branching on attribute *A is:*

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A)$$

It is also the loss function generated by branching on attribute A.

2.3.2. Gini Coefficient

For the CART decision tree, the Gini index is used to select the partitioning attributes. Here, the data set sample set is uniformly set to D. The purity of the data set can be measured by the Gini value:

$$Gini(D) = \sum_{k=1}^{m} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{m} p_k^2.$$

Intuitively, $Gini(D)$ it reflects the probability that two samples randomly selected from data set D have inconsistent class labels. Therefore, $Gini(D)$the smaller it is, the higher the purity of data set D.

Attribute a is defined as:

$$Gini\_index(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$$

Therefore, in the candidate attribute set A, select the attribute that minimizes the Gini coefficient index after partitioning as the optimal partitioning attribute, that is, $a_* = \underset{a \in A}{argmin}\, Gini\_index(D, a)$

2.4. Parameter optimization method

In practical applications, in order to improve the performance and generalization ability of the decision tree model, it is usually necessary to optimize the parameters of the model. The optimization methods mainly include the following main methods. First, the depth of the decision tree determines the complexity and generalization ability of the tree. By adjusting the depth of the tree, you can improve the performance of the model and avoid overfitting or underfitting. Suitable for situations where you need to balance model complexity and performance. You can also control the growth process of the decision tree by adjusting the minimum number of leaf node samples and the minimum number of split samples or the maximum number of features to prevent model overfitting. Then, if there is an imbalanced data set or a noisy data set, the sample weights can be adjusted to balance the training effect of the model and improve the recognition ability of minority class samples. Integrated learning can also be used to combine the prediction results of multiple decision tree models to improve the generalization ability of the model. Common ensemble learning methods include random forests and gradient boosting trees. The parameters of the ensemble

learning method can be adjusted to optimize the performance of the model and improve the accuracy and robustness of the model.

3. Specific application of drug allocation treatment based on decision tree algorithm

3.1. Application background

Drugs data set with a sample size of 200. The patients in the data set all suffer from the same disease. Over the course of treatment, each patient responded to one of drug A, drug B, drug C, drug x, and drug y. This article will build a decision tree model to predict which drug may be suitable for similar patients in the future. The flow chart of the decision tree model for this problem is shown in Figure 2.

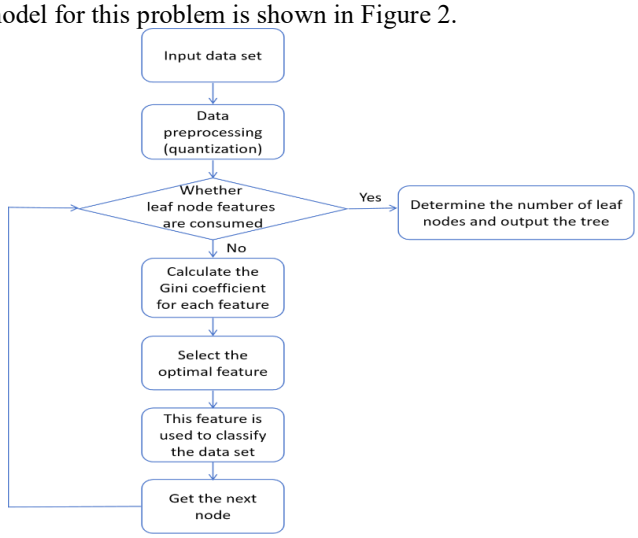


Figure 2 Decision tree model flow chart

3.2. Data preprocessing

Select the sample data of 200 patients. The characteristics of the data set are the patient's age, gender, blood pressure and cholesterol. The goal is to predict each patient 's response to 5 drugs. This is a multi-class classifier sample that uses the training portion of the data set to build a decision tree, which is then used to predict the category of unknown patients, or to prescribe medication for new patients. Part of the data of the sample set is shown in Figure 3.

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	F	HIGH	HIGH	25.355	drugY
1	47	M	LOW	HIGH	13.093	drugC
2	47	M	LOW	HIGH	10.114	drugC
3	28	F	NORMAL	HIGH	7.798	drugX
4	61	F	LOW	HIGH	18.043	drugY

Figure 3 Partial data of the data set sample set

The Drugs data set is a 200*6 matrix, where each row represents a record, and a total of 200 records are sampled. Each column represents a feature of the data set. The structure of the data set is shown in Figure 4.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Age          200 non-null    int64
 1   Sex          200 non-null    object
 2   BP           200 non-null    object
 3   Cholesterol  200 non-null    object
 4   Na_to_K      200 non-null    float64
 5   Drug         200 non-null    object
dtypes: float64(1), int64(1), object(4)
memory usage: 9.5+ KB
None
```

Figure 4 Structure of the data set

Through training, the data set is divided into X and Y categories, corresponding to 134 records and 66 records respectively. Part of the data of the X group training set is shown in Figure 5.

	Age	Sex	BP	Cholesterol	Na_to_K
161	57	F	HIGH	NORMAL	9.945
34	53	M	NORMAL	HIGH	14.133
114	20	F	NORMAL	NORMAL	9.281
94	56	M	LOW	HIGH	15.015
159	34	F	LOW	NORMAL	12.923

Figure 5 Part of the data of X group training set

Quantify the training set data, that is, male = "1", female = "2"; normal blood pressure = "1", high blood pressure = "2"; normal cholesterol = "1", high cholesterol = "2", after quantification Part of the test set data is shown in Figure 6.

	Age	Sex	BP	Cholesterol	Na_to_K
161	57	1	1	1	9.945
34	53	2	2	2	14.133
114	20	1	2	1	9.281
94	56	2	3	2	15.015
159	34	1	3	1	12.923

Figure 6 Part of the data of X group training set after quantification

3.3. Data Visualization and Analysis

This article uses data to draw a visual chart of the paired variable relationship between age and Na to K regarding drugs, as shown in Figure 7.

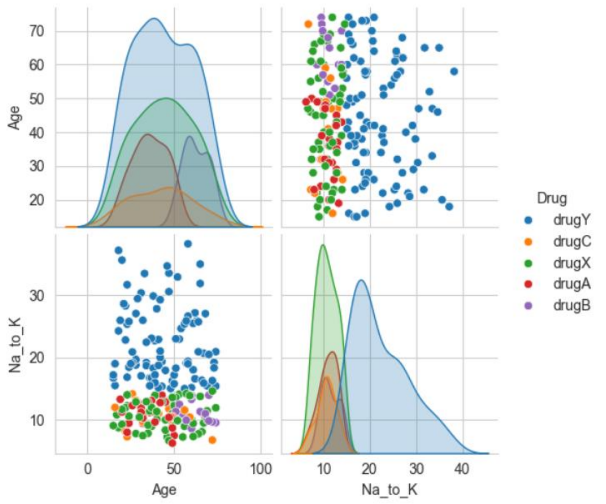


Figure 7 Plot of the relationship between age and Na to K variables on drugs

The decision tree model in this article classifies the training set through the Gini coefficient, sets the maximum depth to 3, and the random index to 0, and obtains the classification results shown in Figure 8.

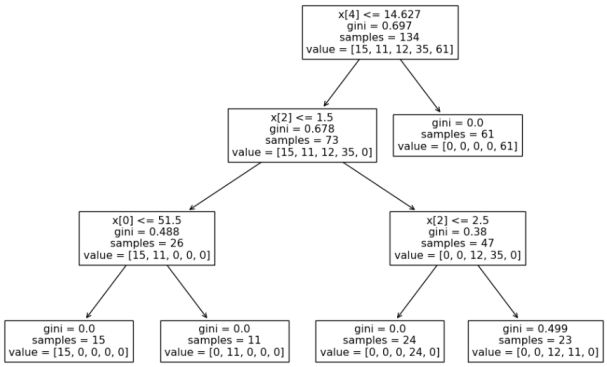
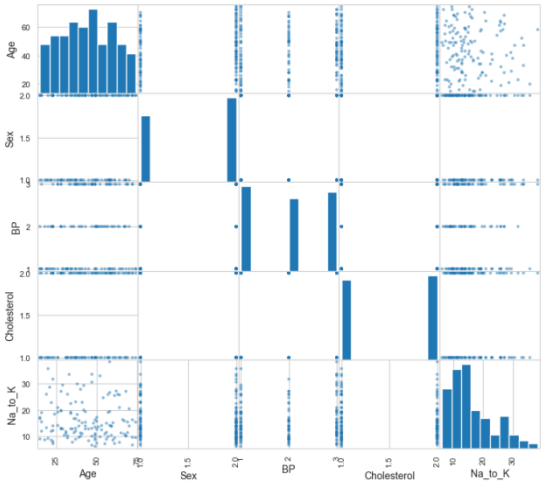


Figure 8 Decision tree classification of Drugs data set

As can be seen from the above figure, the decision tree produces five leaf nodes and five classification results.

The program generates correlation coefficient diagrams and histograms of each feature attribute of the training set and test set, which are used to visually reflect the relationship between feature attributes, as shown in Figure 9 and Figure 10 respectively.



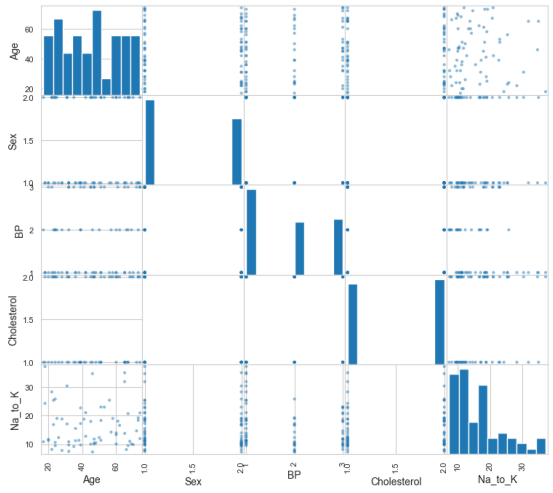


Figure 9 Correlation coefficient diagram of each feature attribute of the training set and test set

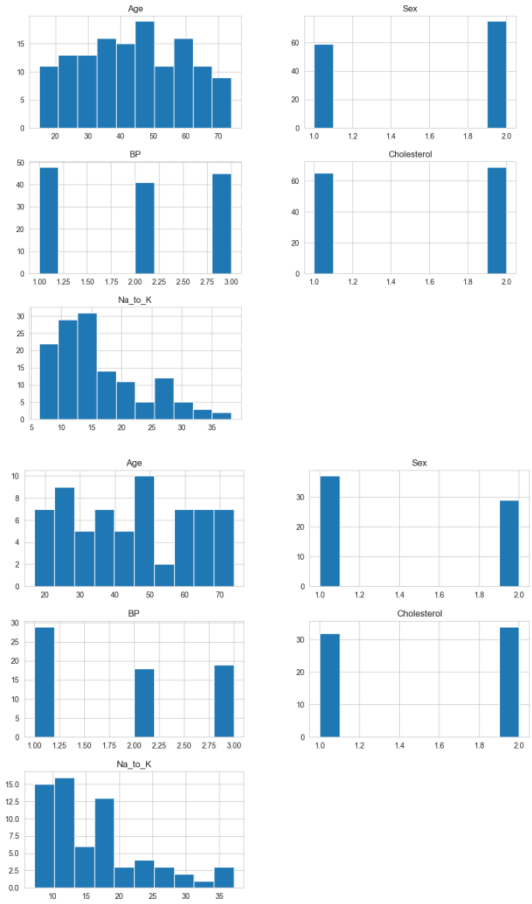


Figure 10 Histogram of feature comparison between training set and test set

As can be seen from the above figure, the characteristic attribute has a strong correlation with itself, but a weak correlation or even no correlation with other characteristic attributes, which shows a high degree of independence of the characteristic attributes.

3.4. Training and classification

After dividing the training set and test set, perform prediction processing on the data set, train and classify the data set, and the results are shown in Figure 11 below.

```
DecisionTreeClassifier(max_depth=3, random_state=0)
['drugX' 'drugX' 'drugY' 'drugX' 'drugY' 'drugY' 'drugY' 'drugB' 'drugX'
 'drugX' 'drugY' 'drugC' 'drugY' 'drugY' 'drugY' 'drugC' 'drugC' 'drugX'
 'drugY' 'drugY' 'drugY' 'drugY' 'drugX' 'drugY' 'drugA' 'drugC' 'drugY'
 'drugY' 'drugY' 'drugC' 'drugA' 'drugY' 'drugY' 'drugY' 'drugB' 'drugC'
 'drugB' 'drugA' 'drugY' 'drugY' 'drugY' 'drugA' 'drugA' 'drugX' 'drugY'
 'drugX' 'drugC' 'drugY' 'drugA' 'drugA' 'drugY' 'drugY' 'drugA' 'drugX'
 'drugX' 'drugC' 'drugB' 'drugY' 'drugC' 'drugY' 'drugY' 'drugC' 'drugA'
 'drugY' 'drugY' 'drugX']
Model accuracy score with criterion gini index: 0.8788
Training-set accuracy score: 0.9179
```

Figure 11 Predicted data set results

As can be seen from Figure 11, the exact value of the model for the standard Gini coefficient is 0.8788, while the exact value of the training set is 0.9179, which shows that the training set can better reflect the situation of the model sample set.

In addition, this application also uses the information entropy method for calculation and comparison, and concludes that the results obtained by the Gini coefficient method and the information entropy method are almost the same in this problem.

4. Conclusion

This article mainly implements the application of decision tree algorithm in drug distribution. First of all, the article introduces the algorithm description of decision tree, depicts the basic idea of decision tree gradually selecting optimal features to divide data, and introduces two algorithm forms of decision tree through information entropy and Gini coefficient, and proposes from them an effective way to optimize parameters. This article mainly uses the decision tree algorithm based on the Gini coefficient to realize the allocation of drugs, and obtains 5 classification results. The number of samples is 61, 15, 11, 24 and 23 respectively, and the standard Gini coefficient corresponding to the predicted value The exact value of the model is 0.8 788, which reflects the good effectiveness of the model for this problem and has more practical significance for research.

References

[1] Cheng Zhongjian. Research on drug-target interaction prediction methods based on deep learning[D]. Central South University, 2024.DOI:10.27661/d.cnki.gzhnu.2022.002640.

[2] FDA.Guidance for industry, investigators and reviewers on exploratory IND studies[EB/OL]. (2006)[2014-03-15]. http://www.fda.gov.

[3] Hunt, Earl B., J Marin and Philip J. Stone. “Experiments in induction.” (1966).

[4] Luo Jia, Li Mingming. Design and application of student employment prediction model based on decision tree algorithm [J]. Integrated Circuit Applications,2023,40(10): 62-64. DOI: 10.19339/j.issn.1674-2583.2023.10.024.

[5] Yu Xiaolin, He Kaiping. Research on disease diagnosis-related grouping of patients with cerebral hemorrhage based on decision tree [J]. Modern Preventive Medicine, 2023, 50(08): 1494-1498+1515.DOI: 10.20043/j.cnki.MPM.202212321.

[6] Feng Yangyi. Travel insurance demand forecast analysis based on decision trees and random forests [J]. Modern Business, 2024(03):130-133.DOI:10.14097/j.cnki.5392/2024.03.019..