Application of Improved Decision Tree C4.5 Algorithms in the Judgment of Diabetes Diagnostic Effectiveness

Jiawei Liu¹, Bo Ning² and Daosheng Shi*¹

¹College of Automation, Wuhan University of Technology, Wuhan 430070, Hubei Province
²College of Automation, Wuhan University of Technology, Wuhan 430070, Hubei Province
*Corresponding author’s e-mail: 2273053784@qq.com

Abstract. C4.5 algorithm needs to scan data sets repeatedly when constructing decision tree, which leads to its inefficient operation. A C4.5 decision tree optimization algorithm based on data dimension reduction is proposed, which generates a more concise decision tree and improves the efficiency of the algorithm. At the same time, it eliminates the interference of redundant attributes and improves the accuracy of the algorithm prediction. The simulation results of the improved C4.5 algorithm and the traditional algorithm show that the accuracy and efficiency of the improved C4.5 algorithm are greatly improved. This algorithm is applied to the prediction system of diabetes diagnosis effect, which provides decision support for the diagnosis of diabetes in hospitals, realizes better precise medical treatment, and makes better use of social medical resources. Successful application of diabetes diagnosis and treatment results also provides technical ideas for large data decision-making in other fields.

1. Introduction

Diabetes mellitus is one of the major diseases threatening human health. How to choose an appropriate and effective treatment plan to control the patient’s condition is an urgent problem to be solved in modern medicine. A therapeutic effect evaluation system is an important means to solve this problem. It is of great significance for the treatment of diabetes mellitus to configure appropriate diagnosis and treatment programs for patients and make full and reasonable use of existing medical resources. In this paper, C4.5 algorithm is used. [1] In order to improve the original model of the evaluation system and reduce the dimension of data, it is more suitable for the application background and more accurate in judging the therapeutic effect.

C4.5 algorithm is an improved ID3 algorithm, which uses information gain rate to select attributes. Decision tree is a greedy algorithm, which follows a unified recursive pattern: (1) Representing a given data set as a root node; (2) Testing a selected attribute on each sub-node after the root node, dividing the data set into subsets, which are represented as a sub-tree; (3) Repeating the above process until the sample in the subset belongs to a category. That is the so-called "pure", so that the growth of the decision tree is completed. Because the classification rules are easy to understand and the prediction accuracy is high, C4.5 algorithm is an important classification algorithm in data mining. However, the algorithm has the following shortcomings: in the process of building decision tree, data sets need to be scanned many times, resulting in inefficiency of the algorithm. [2]
In view of its shortcomings, the paper [3] proposes to introduce Kendall's harmony coefficient into the C4.5 algorithm, which to some extent solves the disadvantage that the original algorithm does not consider the correlation between conditional attributes and improves the accuracy of the algorithm [3]. The paper [4-5] simplifies the calculation of information gain rate in the algorithm and improves the efficiency of the algorithm.[4-5]

In this paper, the feature selection process of data is optimized by using data dimension reduction technology. Data dimensionality reduction is to retain the main information of the data, simplify the data, remove redundant variables in the data set, visualize the data, and improve the efficiency of the algorithm. To a certain extent, it solves the problem of low efficiency of the original algorithm and improves the accuracy of the algorithm.[6]

2. C4.5 algorithm
The C4.5 algorithm belongs to supervised learning. For a given data set, each set of attribute values partitioned represents a tuple, and each tuple is included in a class of classifications in the decision-making process. The purpose of the algorithm is to find some relationship between attributes and categories through training and learning of training set, and then apply this relationship to the examples, classify the new examples and complete the decision-making.

2.1 Selection of partition attributes
Firstly, we complete the following hypothesis: Let A be a tuple training data set of class labels, each attribute has n different values and N different class Bi (i=1,2,...).BiA is a small data set of tuples of category Bi in data set A. The number of meta-ancestors of A and BiA is represented by | A | and | BiA |.

The information gain rate is calculated by using the following formula:
Information entropy:

\[ Info(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \]

Information gain:

\[ Info(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} Info(D_j) \]

\[ Gain(A) = Info(D) - Info_A(D) \]

Information gain rate:

\[ SplitInfo_A(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right) \]

\[ GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \]

The attributes with the largest information gain rate are selected as partition attributes.

2.2 Recursive Process of Decision Tree
Based on the calculated information gain rate, the given data set A is divided into several different non-empty subsets, and the sub-nodes are constructed. Then the sub-nodes grow into sub-trees K and return K. Then the above process is called recursively, and new nodes and subtrees are generated to make the decision tree grow continuously. Up to a certain time, all the samples in the subset belong to a category, and the decision tree does not grow any more.

3. Data Dimension Reduction Technology
If a data has thousands of features, the workload of training the model in such a high-dimensional data set is unimaginable, and the challenge to the model is enormous. At this time, we need a method to help
us to screen variables that are useful to us, reduce the complexity of the training set, and improve the efficiency of algorithm training and the accuracy of the model.

The data dimension reduction technology selected in this paper is PCA. PCA, Principal Component Analysis (PCA), is an unsupervised dimension reduction method that combines high latitude feature vectors with low latitude feature attributes. The objective of this algorithm is to map high-dimensional data to low-dimensional space by linear projection, and to expect the maximum variance (maximum variance theory) of data in projection dimension, so that more original data points can be retained and less data dimension can be used.

Most of the sample information is represented by selecting some variables from the original sample variables. These selected variables are the main components. We can call PCA function to do data principal component analysis.

4. Improvement of C4.5 algorithm
The main idea of improving the original C4.5 algorithm is to use the principal component analysis method to select the principal components of the variables in the data set, to generate the C4.5 decision tree for the simplified data set, and to select several attributes for each node in the principal component analysis method.

The steps of the improved C4.5 decision tree algorithm are as follows:
(1) Data initialization.
(2) Principal component analysis is used to simplify data sets.
(3) Calculating the information gain rate of each principal component
(4) Generate a subset of data, using the principal component with the largest information gain rate as the splitting node.
(5) Repeat steps (3) and (4) so that the decision tree continues to grow until all principal component attributes are utilized.
(6) The decision tree is pruned to generate the decision tree.

5. Analysis of experimental process and results

5.1 Experimental data
The experimental data in this paper come from the database of diabetic patients in several hospitals in the United States, including 32-minute information about patient’s sex, age, medication during treatment, source of admission, place of discharge, and whether to re-enter the hospital. After principal component analysis, several principal components, namely P1, P2, P3, are generated. With a decision attribute, that is, whether to re-hospitalize. Conditional attributes and decision attributes are discrete, so data need not be discretized. The commonly used evaluation system is the binary and ternary evaluation system. In this paper, the binary evaluation system is used to divide the treatment effect into "good" and "general".

5.2 Experimental Process

5.2.1 Principal Component Analysis. The data set in this paper consists of 20,000 samples, which are divided into training set and test set. The training set consists of 12,000 samples, and the training set consists of 8,000 samples. In the training set, the improved C4.5 algorithm is used. First, the principal component analysis method is used to generate the principal component P1-P8, and the results are as shown in Table 1. The information gain rate of each principal component is calculated and the decision tree is constructed according to the information gain rate.
Table 1.

| principal component | characteristic value | Contribution rate | Cumulative contribution rate |
|---------------------|----------------------|-------------------|-----------------------------|
| P1                  | 126.5371             | 75.62%            | 75.62%                      |
| P2                  | 16.03991             | 9.59%             | 85.21%                      |
| P3                  | 9.630419             | 5.76%             | 90.96%                      |
| P4                  | 3.735865             | 2.23%             | 93.19%                      |
| P5                  | 2.712882             | 1.62%             | 94.82%                      |
| P6                  | 1.970874             | 1.18%             | 95.99%                      |
| P7                  | 1.687597             | 1.01%             | 97%                         |
| P8                  | 1.63189              | 0.98%             | 97.98%                      |

5.2.2 Verification of test set prediction results. The decision tree model is applied to the test set, and the accuracy of the original algorithm is stable according to the dimension selection. The accuracy of the improved algorithm is higher than that of the original algorithm, with a maximum of 49.0%. See Figure 1 for details.

Figure 1. Accuracy comparison

5.3 Analysis of experimental results

The decision tree generated by the improved C4.5 decision tree algorithm in this paper is more concise. In this paper, the diagnosis and treatment results are used for leaf nodes, which can obtain the required information more intuitively. When the original algorithm generates the decision tree, all the conditional attributes are used. The improved algorithm only analyses the principal components, establishes decision tree, and gets a faster evaluation. At the same time, because of eliminating the interference of redundant attributes, the accuracy of judgment is higher, and the accuracy rate in the experiment reaches 49.7%. In conclusion, the improved C4.5 algorithm is more concise, effective and superior than the original algorithm in establishing the diabetes diagnosis and treatment system in this paper.

Table 2.

| data set     | Instance number | Attribute number | Accuracy of C4.5 | Accuracy of C4.5 Using PCA |
|--------------|-----------------|------------------|------------------|---------------------------|
| Diabetic_data| 1999            | 31               | 0.4725           | 0.49                      |
6. Conclusions
Medical industry is rich in data, which needs to be analyzed and used to accumulate treatment experience for doctors. Data mining technology can play a huge role in this field. As a decision tree algorithm, C4.5 plays an important role in data mining technology. In this paper, the data dimension reduction technology is used to optimize the process of choosing splitting nodes of C4.5 algorithm, improve the algorithm of C4.5, and apply it to the practical application of diabetes diagnosis effect judgement. The simulation results show that the accuracy of the improved algorithm is higher than that of the original algorithm, and the method has certain advantages over the original C4.5 algorithm.[7]

Acknowledgment
This research is supported by the innovation and entrepreneurship training project in Wuhan University of Technology (Theme of project: Design of reconnaissance system based on STM32 small multi-rotor real-time image transmission, project no.2018-ZDH-B1-09)

References
[1] Quinlan JR. (1993) C4.5: Programs for machine learning. Morgan Kaufmann Publishers Inc. San Francisco.
[2] Zhou ZH. (2016) Machine Learning [M]. Tsinghua University Press. Beijing.
[3] An YP, Shang JZ. (2019) Improvement and analysis of decision tree C4.5 algorithm [J/OL]. Computer engineering and application, 1-6.
[4] Xu P, Lin S. (2009) A traffic classification method based on C 4.5 decision tree [J]. Journal of Software, 20 (10): 2692-2704.
[5] Zhou JF, Yang AM, Liu JC. (2012) Network traffic classification method based on improved C4.5 algorithm [J]. Computer Engineering and Application, 48 (5): 71-74.
[6] Sun PA, Wang BZ. (2019) Research and application of PCA dimensionality reduction method in machine learning [J]. Journal of Hunan University of Technology (01): 80-85.
[7] Yu MM, Zhang JY, Li YE. (2018) Key Technologies and Decision Support for Data Analysis of Medical Health [J]. China Science and Technology Forum, (11): 53-62.