Construction and Application of Prediction Model of Diabetes Treatment Effect Based on Improved CART Algorithm

Construction and Application of Prediction Model of Diabetes Treatment Effect Based on Improved CART Algorithm

Yuchen Qiao, Xu Yang* and Enhong Wu

School of Automation, Wuhan University of Technology, Wuhan, Hubei, China

*Corresponding author's e-mail: yx_auto@whut.edu.cn

Abstract. In this research, the rehospitalisation of diabetic patients was taken as the standard to judge the treatment effect. With the help of CART algorithm improved on the basis of Principal Component Analysis, the researcher explored the relationship between multiple data involved in the diabetic treatment program and its treatment effect and built a prediction model of diabetic treatment effect. The study focused on the treatment data from 130 American hospitals. 9000 sets of data were used as the training set, and 1000 sets of data were used as the test set to generate the decision tree, and then the researcher pruned the generated decision tree. The accuracy of the newly established prediction model was greatly improved by 21% and the running time was largely reduced by 5.352s compared with the old model established on the unimproved CART algorithm. The results of the study well verified the feasibility and high efficiency of the new prediction model established on the improved CART algorithm, and provided a new idea for improving the treatment effect and efficiency of diabetes.

1. Introduction

In recent years, diabetes has become the third major chronic non-communicable disease threatening human health after cancer and cardiovascular disease, and it is an increasingly serious public health problem [1]. Currently, there are about 150 million people with diabetes worldwide, which is expected to increase to 300 million by 2025. The grim situation of diabetes in China is even more sobering. By 2003, China had become the second largest country with diabetes. Therefore, it is of great practical significance to study the factors affecting the treatment of diabetes, establish a prediction model for the treatment effect of diabetes, and improve the treatment effect and efficiency of diabetes.

At present, many studies have been carried out at home and abroad on factors affecting the therapeutic effect of diabetes, including gender, age of patients and various drugs in doctors' treatment programs, etc., but there are still few studies on the prediction model of therapeutic effect of diabetes. The modelling methods to solve this kind of problem include linear analysis, multivariable linear regression and neural network nonlinear analysis. However, due to the large number of variables affecting the therapeutic effect of diabetes and the non-linear relationship between variables and therapeutic effect, the accuracy of predicting therapeutic effect of the first two methods is low. Although the third method can be analysed from a more fitting point of view, it is difficult to control the parameter setting of neural network for different data. Therefore, this research used the method of data mining to establish the prediction model of therapeutic effect. Compared with other methods, it can process large-scale data efficiently and build models. The decision tree algorithm in data mining has the advantages of simple calculation and easy operation, so this research adopted the CART algorithm in data mining to study the relationship between factors affecting the therapeutic effect of diabetes and the final therapeutic result.

Since the impact factors of diabetes treatment effect are complex and large in quantity, this study adopted principal component analysis in data processing. Firstly, all impact factors were simplified into several main ones, and then the data mining with CART algorithm was carried out, which greatly improved the efficiency of data processing. In this paper, the decision tree generated by CART algorithm was pruned to avoid too many branches of the decision tree. The fitting of training data was perfect, but the fitting of test data was poor. By pruning, the risk of overfitting was reduced and the efficiency of the prediction model for the therapeutic effect of diabetes was improved.

2. Materials and methods

2.1. Materials

The research targeted at the treatment data of diabetes from 130 American hospitals. The incomplete data and the irrelevant variables such as encounter_id and patient_nbr were all removed. The impact factors we chose to study included gender, age, admission_type_id, discharge_disposition_id, admission_source_id, time_in_hospital, change, diabetesMed, readmitted and 23 kinds of drugs.

In this research, the readmission time of diabetic patients after treatment was taken as the basis in judging the treatment effect. The treatment effect was considered as good if patients received no readmission treatment or readmission happened 30 days later, while the treatment effect was considered as poor if patients got readmitted within 30 days. This judgment is linked with reality and easy to implement, so it can judge the treatment effect more precisely.

2.2. Research method

2.2.1. Principle Component Analysis. Principal Component Analysis (PCA) uses the idea of dimensionality reduction, which is a method to recombine many original indicators with certain correlation into a new set of a few unrelated comprehensive indicators [2]. That is:

$$Fp = a_{1i} * Z_{Xp} + a_{2i} * Z_{Xp} + ... + a_{pi} * Z_{Xp}, \quad i = 1, ..., p \tag{1}$$

The a1i,…, api (i = 1,…, p) is the eigenvectors corresponding to the eigenvalue of Σ -- the covariance matrix of X. ZX1, ZX2 …, ZXp is the standardized value of the original variable.

We first used SPSS software to standardize the original data, and then carried out correlation analysis between the indicators. After determining the number of principal components, we calculated the new expression of principal components. By means of PCA, we reduced the dimensions of 32 factors influencing the diabetes treatment results to simplify the structure of the decision tree generated by CART algorithm.

2.2.2. CART algorithm. Decision tree algorithm is a typical classification method to approximate discrete function in data mining. After processing the data, the induction algorithm is used to generate the decision tree, and then the decision is used to analyse the new data. The goal is to extract decision steps, rules, patterns, and knowledge from archived databases [3].

In the decision tree algorithm, CART (Classification and Regression Tree) algorithm is a very effective non-parametric Classification and Regression method [4]. This algorithm is a non-parametric statistical method used to classify discrete or continuous dependent variables. Its basic principle is to form a decision tree structure in the form of binary tree through the cyclic analysis of the training data set composed of test variables and target variables [5].

CART algorithm uses binary recursion to divide the current data set into two subsets [6]. When the feature attribute is a continuous value, a certain feature attribute, ai, should be selected, and then a value bi of the feature attribute should be selected, which divides the space into ai≤bi and ai>bi. Then recursively divide the two parts in this method until the data set is completely divided.

CART algorithm uses the "Gini Index" to measure the impurity of the split. For a node t of the decision tree, its Gini index is calculated as follows:

$$Gini(t) = 1 - \sum_{k} [p(ck \mid t)]^2 \qquad (2)$$

The Gini index is the difference between 1 and the sum of the probability squares of category Ck, reflecting the degree of uncertainty of the sample set. Gini coefficient measures the impurity of sample division or training sample set. The smaller the impurity, the higher the "purity" of the sample [7], and the lower the degree of uncertainty of the sample set. Therefore, in the process of division, the characteristic splitting of the minimum Gini index should be selected.

Suppose that the sample set corresponding to the parent node is A when splitting, and feature B is selected by the CART algorithm to split into two child nodes, corresponding to the set AL and AR, then the formula for calculating the Gini index after splitting is as follows:

$$Gini(A,B) = \frac{|A_L|}{|A|} Gini(A_L) + \frac{|A_R|}{|A|} Gini(A_R) \qquad (3)$$

In the process of constructing the classification decision tree of the diabetes treatment effect prediction model, the Gini coefficient of the nodes was calculated according to the above formula, and the optimal segmentation threshold and the optimal test variables were selected according to it to divide the nodes, and the optimal decision tree was finally generated recursively.

3. Construction and application of the prediction model of diabetes treatment effect based on improved CART algorithm

The traditional CART algorithm is more complex in the process of discrete classification of continuous attributes, and requires a large amount of computation. Meanwhile, the prediction accuracy of small sample data is relatively low [8]. This research first adopted PCA to analyse the impact factors and reduce the dimension; and then studied the relationship between the influence factors and diabetes treatment effect by means of CART decision tree algorithm. Meanwhile, the CART decision tree algorithm was optimized and pruned so as not to affect the accuracy of the prediction model because of overfitting.

3.1. Results of principal component analysis

The 32 input variables caused by Data concentration and influencing treatment effect include gender, age, admission_type_id, discharge_disposition_id, admission_source_id, time_in_hospital, change, diabetesMed, readmitted and 23 kinds of drugs. We use MATLAB programming to conduct principal component analysis and the eigenvalues and variance contribution rates of the correlation coefficient matrix R in the sample data set are shown in table 1.

Table 1.the eigenvalues and variance contribution rates of R

The principal components	The eigenvalue	Variance contribution rate	Cumulative contribution rate
F1	28.2883	0.471	0.471
F2	16.5389	0.2753	0.7463
F3	8.6441	0.1439	0.8902
F4	2.5989	0.0433	0.9335
F5	2.0478	0.0341	0.9676

In practical applications, in order to make full use of the original information, the corresponding components whose cumulative contribution rate is above 85% are generally selected as the retention components [9]. As can be seen from table 1, the contribution rate of the first five principal components has reached 96.76%, indicating that the first five principal components basically contain all the information of the features. Therefore, we used the first five principal components and the original data to form a new sample set, and we named the new variables F1, F2, F3, F4 and F5.

3.2. Model establishment and application of CART decision tree algorithm based on principal component analysis

3.2.1. Model establishment of CART decision tree algorithm based on principal component analysis. The process of establishing the model of CART decision tree algorithm based on principal component analysis is as follows:

- Obtain 10,000 sets of data after dimension reduction by main analytic hierarchy process.
- Use data set to calculate the index representing the degree of therapeutic effect and judge the therapeutic effect.
- Select part of data in the data set randomly as the training set, take five principal component variables as input variables, and the treatment effect as the target variable, and use Matlab software to build the model with CART algorithm.
- Prune the obtained CART algorithm decision tree.
- Select the remaining data in the data set as the test set to verify the accuracy of the model.

3.2.2. Model application of CART decision tree algorithm based on principal component analysis. Based on the treatment data of 10,000 patients with diabetes from 130 American hospitals, a prediction model of the treatment effect of diabetes was established. We took 9000 groups as the training set and used CART algorithm to calculate the five main impact factors to get the decision tree, as shown in figure 1.

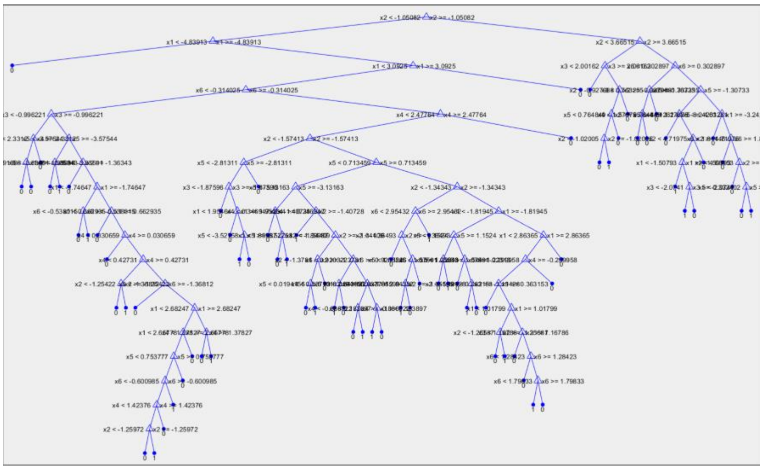

Figure 1. Decision tree of therapeutic effect prediction generated by improved CART algorithm

Then the optimal pruning was carried out for the decision tree to obtain the decision tree after pruning, as shown in figure 2.

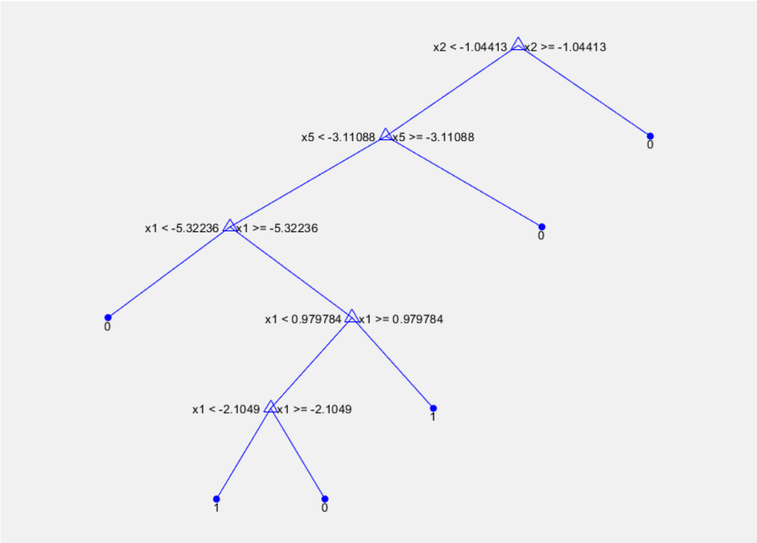

Figure 2. decision tree for predicting therapeutic effects after pruning

Finally, we took the remaining 1000 sets of data as the test set, and the accuracy of the prediction model generated by the improved CART algorithm was 73%.

4. Comparison between improved CART algorithm prediction model and CART algorithm prediction model

We adopted CART decision tree algorithm to build a diabetes treatment effect prediction model, and compared it with the model established above, so as to draw a conclusion more intuitively. Similarly, 9000 sets of data were used as the training set to construct the decision tree with CART algorithm, as shown in figure 3.

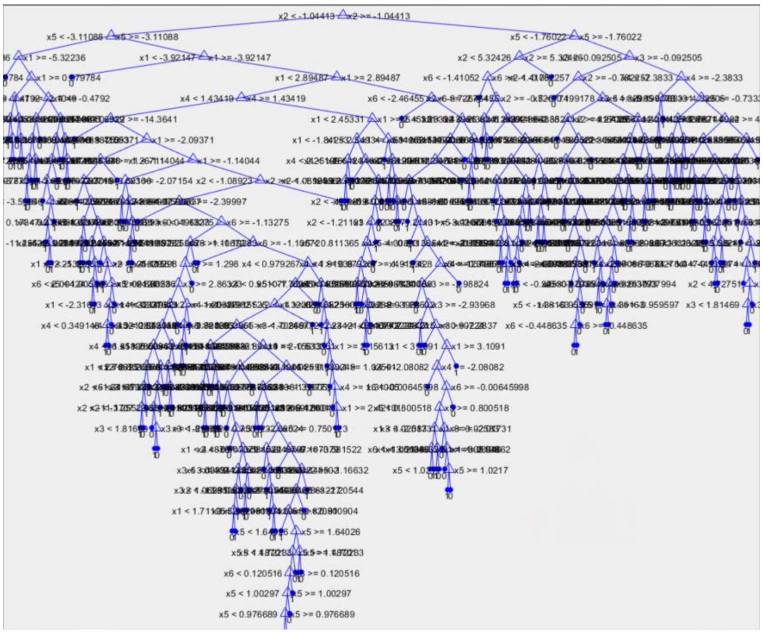

Figure 3. Decision tree of therapeutic effect prediction generated by CART algorithm

We compared the decision tree generated before and after improving CART algorithm and it was clearly shown that that the decision tree generated by the improved CART algorithm was more concise. The remaining 1000 sets of data were taken as the test set, and the accuracy rate of CART algorithm before improvement was 52%.

The running time and prediction accuracy of the model obtained by experiments were compared with the traditional CART decision tree algorithm, and the results are shown in table 2.

Table 2.Comparison of predicted results

Algorithm	The prediction accuracy of diabetes treatment effect /%	The elapsed time /s
Traditional CART algorithm	52.00	15.3496
Improved CART algorithm	73.00	9.9976

It can be clearly seen from table 2 that the prediction accuracy of the model increased by 21%, from 52.00% to 73.00%. The expected time was reduced by 5.352s, from 15.3496s to 9.9976s. The accuracy rate was greatly increased and the running time was greatly reduced. It can be concluded that the improved CART decision tree model has a very good applicability in predicting the therapeutic effect of diabetes.

5. Conclusion

There is a large research blank in making use of machine learning algorithm to predict the therapeutic effect of diabetes. Due to the large number of factors affecting the therapeutic effect of diabetes, and the complicated and non-linear relationship between the factors and the therapeutic effect, the establishment of the prediction model is greatly hindered. To solve the above problems, we first adopted PCA to reduce the dimension and complexity of the original 32 impact factors. And then based on the treatment data of 10,000 diabetic patients from 130 American hospitals, we built a new prediction model for diabetes treatment effect by means of improved CART algorithm. With experimental comparison, it was proved that the accuracy of the model improved from 52% to 73% and the running time reduced from 15.3496s to 9.9976s, which showed that the model built in this research had a good reliability.

The accuracy of the prediction model built by means of data mining method can be further improved. However, the combination of data mining method and diabetes treatment can help make clear the relationship between the diabetic impact factors and the diabetic treatment effect, which can not only make data mining method more extensively adopted but also provide a new thought in improving the treatment efficiency of diabetes.

Acknowledgements

This paper is supported by National Students' Platform for Innovation and Entrepreneurship Training Program in 2018(ID:201810497222)

References

[1] Zhu, G.,(2006) Research progress in health education for diabetes patients (review).J. China urban and rural enterprise health.,2006(06):57-60.

[2] Zhang,S., Zhang, C., Meng,Q.(2013) Comprehensive evaluation of regional technological innovation capacity in China based on principal component analysis .J. Economic journal.,2013(Z2):90-91.

[3] Friedl, M.A.， Brodley,C. E.(1997)Decision tree classification of land cover from remotely sensed data.J.Remote Sens． Environ.,61(3):399-409．

[4] Chen, H., Xia, D.(2011) Research on application of data mining algorithm based on CART decision tree .J. Coal technology.,2011,30(10):164-166.

[5] Qi, L., Yue, C.(2011) Classification of remote sensing images based on CART decision tree method .J. Forestry survey planning.,2011,36(02):62-66.

[6] Breiman, L.,Friedman, J. H.,Olshen, R. A.,et al.(1984)Classification and regression trees (CART) .J. Encyclopedia of Ecology,1984,40(3):582-588.

[7] Liu, C.(2013) Cost sensitive decision tree generation method based on relational degree .J. Journal of changchun university of technology (natural science edition).,2013,34(02):218-222.

[8] Tang, L., Li, L.(2008) Knowledge acquisition prediction in m-learning process based on improved CART algorithm .J. Journal of shaoguan university., 2008,39(09):26-31.

[9] Tian,Y., Ma,W., Liu,Y., Xiao, Z.,Cheng, G.(2014) Prediction of water flooding effect in oil Wells based on pca-fnn.J. Journal of xi 'an petroleum university (natural science edition). **29(03)** 83–86+10–11