WEP Warwick Evans Publishing

Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 4
ICCIA 2024

Research and Application of Random Forest Model Based on Genetic Algorithm Optimisation

Lixiang Yu, Wenkui Wu*

School of Mathematical Sciences, Huaqiao University, Quanzhou, China

*Corresponding author: wwkwwk1626@163.com

Abstract. In this study, an in-depth analysis of the relationship between infant behavioural characteristics and mothers' physical and psychological indicators was conducted by integrating a random forest model optimised by a genetic algorithm. A mathematical model of treatment cost and health improvement rate was also established, which provides a scientific basis and technical support for infant behaviour analysis and individualized intervention strategies.

Keywords: Random Forest, Prediction Model, Genetic Algorithm, Big Data.

1. Introduction

The mother is one of the most important people in the life of the infant, providing not only nutrients and physical protection, but also emotional support and a sense of security. Poor mental health status of the mother, such as depression, anxiety and stress, may have a negative impact on the cognitive, emotional and social behaviour of the infant. Stressed mothers may negatively affect the physical and psychological development of their infants, for example in terms of sleep.

2. Data sources and pre-processing

The data used in this paper comes from the Huashu Cup Mathematical Modelling Competition C question. In the data pre-processing, this paper adopts filling the outliers into the plural to deal with; the sleep time of 99:00 for the whole night belongs to the obvious error, which is excluded to exclude the influence of the outliers on the data; for the sleep time of the whole night, it is enough to directly convert the time format into the numerical format.

3. Introduction to the methodology

3.1. Random Forest

The Random Forest algorithm is a machine learning algorithm based on the idea of integrated learning for classification and regression tasks [1]. The basic idea of the algorithm is to construct multiple decision trees and combine their predictions into voting or weighted voting to get the final prediction [2]. Specifically, the Random Forest algorithm first randomly selects a certain number of samples from the original training data set, then randomly selects a certain number of features from these samples, and constructs multiple decision trees based on these features. In the process of constructing decision trees, Bootstrap sampling and random selection of features are used to make each decision tree have a certain degree of independence and difference [3].

In the classification task, for a new unknown sample, each decision tree predicts its classification, and selects the category with the highest number of occurrences as the category of the unknown sample according to the principle of majority voting. In the regression task, each decision tree predicts the regression of a new unknown sample and calculates the final regression result based on the weighted average method. The Random Forest algorithm has high classification accuracy and low risk of overfitting, and is suitable for dealing with high-dimensional data and unbalanced data. In

practical application, some optimisation techniques and feature selection methods are often used to improve the performance and accuracy of the algorithm.

3.2. Genetic Algorithm

Genetic algorithm is an optimisation algorithm based on the principle of biological evolution, which can be used to search for optimal solutions [4]. In random forest algorithms, genetic algorithms can be used to optimise the parameters of the algorithm to obtain better classification or regression results. Specifically, genetic algorithms treat the parameters of the Random Forest algorithm as biological genes, and by simulating the process of biological evolution, they constantly mutate, crossover and select the parameters in order to find the optimal combination of parameters. In this process, the genetic algorithm can automatically search and adjust the parameters of the Random Forest algorithm, such as the number of decision trees, the number of features, the size of subsets, etc., in order to obtain the best model performance [5].

Compared with the traditional grid search and random search methods, genetic algorithms can search for the optimal solution more comprehensively and quickly when optimising the parameters of the Random Forest algorithm, and can handle more combinations of parameters. Therefore, genetic algorithms play an important role in parameter optimisation of the Random Forest algorithm to improve the accuracy and efficiency of the model [6].

4. Genetic Algorithm Parameters for Optimisation

The parameters of our prescribed genetic algorithm in SPSSPRO are as follows in Table 1:

Table 1. Genetic algorithm parameters

Parameter	Numerical Value
Initial number of populations	50
Maximum number of iterations	100
Mutation probability	0.01
Crossover probability	0.5

A random number of optimisation parameters were obtained as shown in the Table 2 below:

Table 2. Random forest optimisation parameters

Parameter Name	Parameter Value
Data slicing	0.7
Data shuffling	yes
Cross-validation	10
Node splitting evaluation criteria	gini
Number of decision trees	1000
With playback sampling	true
Out-of-bag data testing	true
Maximum proportion of features considered for splitting	auto
Minimum number of samples for internal node splits	2
Minimum number of samples in leaf nodes	1
Minimum weight of samples in leaf nodes	0.011
Maximum depth of the tree	23
Maximum number of leaf nodes	50
Threshold value for impurity of node division	0

Then the accuracy of the training set, cross-check set and test set of the random forest model trained here is generally good, and the specific values are shown in Table 3:

Table 3. Random Forest Prediction Accuracy

	Accuracy	Recall	Precision
Training set	0.906	0.906	0.916
Cross validation set	0.534	0.534	0.460
Test set	0.6	0.6	0.494

The fact that the accuracy of the test set did not reach more than 0.7 or even 0.8 is in fact explicable; the data for the eight categories of the independent variable are then rather redundant, and the eight indicators for mothers correlate with the behavioural traits of the infants, but do not appear to be stronger, and thus a bottleneck is encountered in the prediction here.

5. Treatment programme cost model

5.1. Establishment of treatment cost equations

According to 'the rate of change of treatment costs relative to the degree of illness are proportional to the cost of treatment' can be derived from the differential equation (1), and then its transformation into the equation (2) after the integration of the two sides of the equation and then derived from equations (3) and (4), and the equation (4) that is the form of analytical solution of the original equation. An expression for the rate of change of the treatment cost with respect to the three types of psychological indicators is established. The CBTS scores were used as the x-axis coordinates, the EPDS scores were used as the y-axis coordinates, and the HADS scores were used as the z-axis coordinates.

$$\frac{dy}{dx} = ky \tag{1}$$

$$\frac{dy}{y} = kdx \tag{2}$$

$$\ln y = kx + c \tag{3}$$

$$y = e^{kx+c} \tag{4}$$

w_1 is the cost of treatment for CBTS, w_2 is the cost of treatment for EPDS, and w_3 is the cost of treatment for HADS. and w is the total treatment cost. It is known that the form of the expression is similar to (4), and the following result can be obtained:

$$w_1 = 200e^{0.881x} \tag{5}$$

$$w_2 = 500e^{0.665y} \tag{6}$$

$$w_3 = 300e^{0.746z} \tag{7}$$

$$w = w_1 + w_2 + w_3 \tag{8}$$

5.2. Constructing a three-dimensional spatial coordinate system

Before finding the best treatment plan, we should observe the scatter points in the three-dimensional spatial coordinate system, mainly observe the distribution of scatter points of the infants' behavioural characteristics of quiet, moderate and ambivalent, and use MATLAB to draw a scatter plot as in Figure 1.

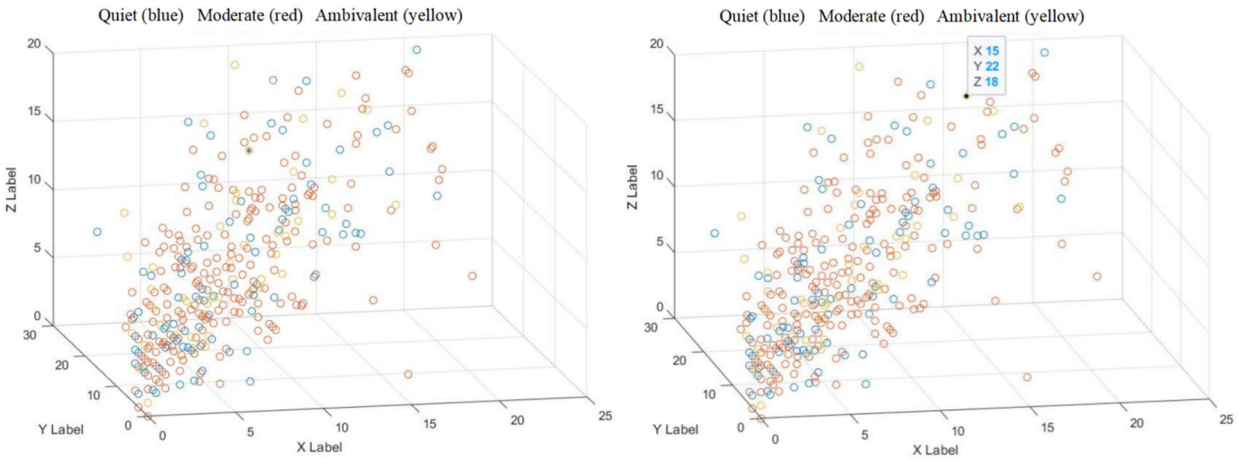

Figure 1. Sample Scatter Plot and Sample Location 238

It is clear that the scatters of Quiet, Medium and Ambivalent are distributed haphazardly, but they are all concentrated around the diagonal of the cube, but it is difficult to visualise the relative distribution of the scatters of the three types of infant behavioural traits. It is difficult to see the relative distribution pattern among the three types of infant behavioural traits. There are also cases where multiple scatters of both types occupy the same position at the same point in the space. If we use the general clustering method, it may not be able to describe the distribution of scatters well. Therefore, we choose to construct an approximate k-nearest neighbour screening algorithm based on sample density and Manhattan distance to find the best solution. It is not appropriate to increase the three psychological indicators for mothers, as an increase in their values would represent an aggravation of the pathology. This means that the best point to look for is in the three-dimensional space defined by the coordinates of point 238 and the origin of the coordinates. In other words, the horizontal, vertical, and orthogonal coordinates of the point must not be greater than the horizontal, vertical, and orthogonal coordinates of point 238, and the three coordinates of the scattering point must be non-negative integers. Using Python to eliminate the scatters outside the constraints and the contradictory points within the constraints, and further counting the quiet and medium points within the constraints, we can get that the quiet type accounts for about 35% of the points. Provision: If the percentage of quiet points in the 6 known scatter points nearest to the Manhattan is more than 35%, the point is considered as 'Strictly Quiet'; if the percentage is less than 35%, the point is considered as 'Strictly Medium'.

6. Conclusions

In this paper, the relationship between the behavioural characteristics of the infants and the physical and psychological indicators of the mothers is modelled, and a genetic algorithm is used to optimise the parameter values of the random forest to arrive at the best prediction accuracy. In the end, out of the 20 groups to be predicted, there were 3 infants with a quiet behavioural profile and 17 infants with a moderate behavioural profile. An expression for the cost of treatment with respect to the rate of change in the degree of illness was solved to create an equation, and an approximate k-nearest neighbour screening algorithm based on sample density and Manhattan distance was established using the three psychological indicators as the coordinates of the x,y, and z axes, and a plausible scenario for switching from Ambivalent 238 to Moderate, as well as Quiet, was given.

References

[1] Naghibi S A, Ahmadi K, Daneshi A. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping[J]. Water Resources Management, 2017, 31: 2761-2775.

[2] Elyan E, Gaber M M. A genetic algorithm approach to optimising random forests applied to class engineered data[J]. Information sciences, 2017, 384: 220-234.

[3] Ansótegui C, Malitsky Y, Samulowitz H, et al. Model-Based Genetic Algorithms for Algorithm Configuration[C]//IJCAI. 2015: 733-739.

[4] Cerrada M, Zurita G, Cabrera D, et al. Fault diagnosis in spur gears based on genetic algorithm and random forest[J]. Mechanical Systems and Signal Processing, 2016, 70: 87-103.

[5] Wang H, Jin Y. A random forest-assisted evolutionary algorithm for data-driven constrained multiobjective combinatorial optimization of trauma systems[J]. IEEE transactions on cybernetics, 2018, 50(2): 536-549.

[6] Zhu E, Chen Z, Cui J, et al. MOE/RF: a novel phishing detection model based on revised multiobjective evolution optimization algorithm and random forest[J]. IEEE Transactions on Network and Service Management, 2022, 19(4): 4461-4478.