Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/75/20240532

Particle swarm optimization random forest based data regression prediction algorithm for melanoma size prediction

Zhanyuan Chen[1,*], Leyao Li[2]

[1]College Of Management Science, Chengdu University of Technology, Sichuan, Chengdu, 610059, China
[2]College Of Management Science, Chengdu University of Technology, Sichuan, Chengdu, 610059, China

[*]Corresponding author email: 342368103@qq.com

Abstract. In this study, a data regression prediction algorithm based on particle swarm optimization of random forests is dedicated to the accurate prediction of melanoma size. As a malignant skin tumor, melanoma size prediction is crucial for treatment and intervention. By dividing the dataset into training and test sets in the ratio of 7:3, we found that in the training set, the model's prediction of melanoma size was highly consistent with the actual value, indicating that the proposed algorithm has good prediction performance. Meanwhile, in the test set, the predicted values of melanoma size are also close to the actual values, verifying the robustness and generalization ability of the model. Specifically, the mean absolute error (MAE) was 1.5682 in the training set and 4.058 in the test set.This indicates that our model can accurately predict the melanoma size and can be further optimized to improve the prediction accuracy in the future. Overall, the data regression prediction algorithm based on particle swarm optimization of random forests provides new ideas and methods for accurate prediction of melanoma size and is expected to play an important role in clinical practice. This research result lays a solid foundation for further exploration and application in melanoma-related fields.

Keywords: Random forest, Melanoma size prediction, Data regression.

1. Introduction

Melanoma is a malignant skin tumor that develops from melanocytes and usually appears on the surface of the skin, but may also arise in other areas [1]. Melanomas come in various shapes, sizes and colors and may not be readily apparent in the early stages but may become visible as the disease worsens. The main reason for this is exposure to ultraviolet light, which is especially dangerous for people with low levels of skin pigment melanin [2,3].

In recent years, with the wide application of machine learning algorithms in the medical field, predicting melanoma size has become a research topic of great interest. Machine learning algorithms can help doctors predict melanoma size more accurately and provide personalized treatment plans by analyzing and learning from a large amount of patient data.

Machine learning algorithms play an important role in the context of melanoma research. Traditional clinical diagnosis mainly relies on doctors' experience and expertise, but it is often subjective and limited [4,5]. In contrast, machine learning algorithms can improve the prediction accuracy and treatment effect

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/75/20240532

by deeply analyzing and mining a large amount of patient data and discovering the patterns and features hidden in the data. Specifically, in terms of predicting the size of melanoma, machine learning algorithms can analyze multidimensional information such as patients' clinical data and imaging data by building complex mathematical models, and make a comprehensive assessment by combining multiple factors such as genetic factors and environmental factors [6]. This personalized prediction model can help doctors better understand the development trend of the patient's condition and adjust the treatment plan in time.

Machine learning algorithms are of great significance in predicting the size of melanoma, which can provide clinicians with scientific basis and decision-making support, and help to improve the therapeutic effect and survival rate of melanoma patients. In this paper, the data regression prediction algorithm based on particle swarm optimization random forest predicts the size of melanoma, which provides a new method in the research of melanoma size prediction. With the continuous progress of technology and data accumulation, it is believed that machine learning algorithms will play an increasingly important role in the field of melanoma in the future.

2. Source of data sets

This paper uses the open-source dataset on the Kaggle website for the study, the data URL is (ht tps://www.kaggle.com/datasets/anmolkumar/machine-hack-melanoma-tumor-size-prediction?rvi=1),th -e current data citation The volume is 14, and the overall characteristics of the data are intact wit h no missing data. The data include information on various patient dimensions, such as the qualit y of melanoma tumor area replacements, the size of the melanoma tumor study area, the ratio of normal to malignant surface replacements, the area of skin damaged by the tumor that cannot be recovered, the total area of skin exposed to the tumor, the standard deviation of the malignant ski n measurements, the ratio of malignant skin measurement error damage to the total spread of the skin, and the size of the melanoma, where melanoma size is the target variable to be predicted, a nd the size of the skin is the target variable to be predicted, and melanoma size is the target vari able to be predicted. size was the target variable to be predicted and the data dimension was 330 × 10.

3. Method

3.1. Particle swarm algorithm

Particle swarm algorithm is an optimization algorithm based on group intelligence, inspired by the behavior of groups of organisms, such as flocks of birds or schools of fish, during their search for food or migration. The particle swarm algorithm was first proposed by Kennedy and Eberhart in 1995, and its basic idea is to search for the optimal solution by simulating the collaborative and competitive relationship between individuals in a flock of birds.The PSO algorithm has the advantages of simplicity, ease of implementation, and does not require gradient information, and has been widely used in the fields of function optimization, neural network training, combinatorial optimization, and so on.

In particle swarm algorithm, the solution space of the problem to be solved is regarded as a certain point in a multidimensional space, which is called a particle. Each particle has its own position and velocity, and updates its position and velocity according to its own experience and the experience of its neighboring particles. By continuously adjusting its position, the particle gradually approaches the global optimal solution, thus realizing the optimization of the objective function.

The core idea of particle swarm algorithm is to guide the particle search by introducing two important factors: individual optimization and global optimization. Each particle remembers the best position it has found in its own history, as well as the best position it has found in the entire population. When updating the position, the particle will be adjusted based on these two pieces of information, taking into account both individual experience and group collaboration to achieve a more efficient search process.

The basic steps of the particle swarm algorithm include initializing the particle swarm, evaluating the fitness, updating the individual optimum and global optimum, and updating the speed and position.

DOI: 10.54254/2755-2721/75/20240532

In practical application, the parameters need to be adjusted according to specific problems, such as learning factor, inertia weight, etc. Meanwhile, when dealing with complex problems, hybrid optimization can also be combined with other heuristic algorithms or improvement strategies to improve the convergence speed and global search capability.

3.2. Random forest algorithm

The Random Forest algorithm is an integrated learning method, proposed by Leo Breiman and Adele Cutler in 2001. It performs classification or regression tasks by constructing multiple decision trees and combining their predictions. Random forest algorithm performs well in solving complex machine learning problems with high accuracy, strong robustness and not easy to overfitting, so it is widely used in data mining, pattern recognition, bioinformatics and other fields. The model diagram of random forest is shown in Figure 1.

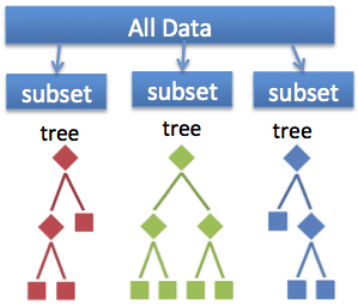

Figure 1. The model diagram of random forest.
(Photo credit : Original)

The core idea of the Random Forest algorithm is integrated learning by constructing multiple decision trees. In the training process of each decision tree, self-sampling will be used to take samples from the original training set with put back, and then the decision tree will be constructed using feature random selection. This ensures the variability between each decision tree, avoids overfitting, and improves the generalization ability of the overall model [7].

The random forest algorithm considers multiple features for segmentation at each node, rather than selecting only one best feature [8].

3.3. Random Forest Regression Algorithm Based on Particle Swarm Optimization

The combination of PSO and Random Forest can improve the parameter optimization efficiency and prediction performance of Random Forest regression models [9]. In this combination, the PSO algorithm is used to search for the optimal combination of hyperparameters, i.e., the number of decision trees, the maximum depth, the minimum number of samples in leaf nodes, etc. By continuously adjusting these hyper-parameters and combining them with methods such as cross-validation for model evaluation, the optimal combination of parameters can be found to improve the model performance [10].

The random forest regression algorithm optimized by particle swarm algorithm has wide applicability and good results in practical applications. Through PSO optimization, the optimal parameter combination can be found faster, and the prediction accuracy can be improved while guaranteeing the generalization ability of the model. This combination method is not only applicable to regression problems, but also has similar application scenarios in classification problems.

4. Experimental setup

Before conducting the experiment, it is necessary to divide the dataset and set the basic parameters of the model used in the experiment.The computer used in this experiment is a 2080 graphics card with 32G RAM. In the data division, the data set is divided into training set and test set according to the ratio

DOI: 10.54254/2755-2721/75/20240532

of 7:3. The training set is used for model training, and the test set is used to evaluate the model performance. The specific parameter settings of the particle swarm algorithm are shown in Table 1, and the specific parameter settings of the random forest algorithm are shown in Table 2.

Table 1. Particle swarm algorithm parameterization.

Particle swarm algorithm parameterization	Parameters
Number of particles	50
Maximum number of iterations	100
Inertial weighting	0.7
learning factor	2

Table 2. Parameterization of random forest regression models.

Parameterization of random forest regression models	Parameters
Number of decision trees	100
Maximum number of features	sqrt
Maximum tree depth	None
Sample size of the smallest leaf node	1

For the particle swarm algorithm parameter settings, the number of particles is set to 50 particles, which represents the number of sample points in the search space. The maximum number of iterations is set to 100 iterations to control the maximum number of times the algorithm runs. The inertia weight is set to 0.7 to balance the trade-off between local and global search. The learning factor is set to 2.0 to control the learning speed during the particle velocity update process.

In the random forest regression model parameter settings, the number of decision trees is set to 100 decision trees to control the number of base learners in the random forest. The maximum number of features is set to sqrt(total number of features), which is the maximum number of features considered in splitting nodes for each decision tree. The maximum tree depth is set to None to avoid overfitting. The minimum leaf node sample number is set to 1 to ensure that there are enough samples on the leaf nodes.

Through the above parameter settings, the Random Forest regression model is tuned and optimized in conjunction with the particle swarm algorithm, and better performance can be obtained in the melanoma size prediction task.

5. Result

After the model training is completed, the feature importance parameters are output, as shown in Figure 2, along with a line graph of predicted-actual values for the training set and predicted-actual values for the test set, as shown in Figures 3 and 4.

DOI: 10.54254/2755-2721/75/20240532

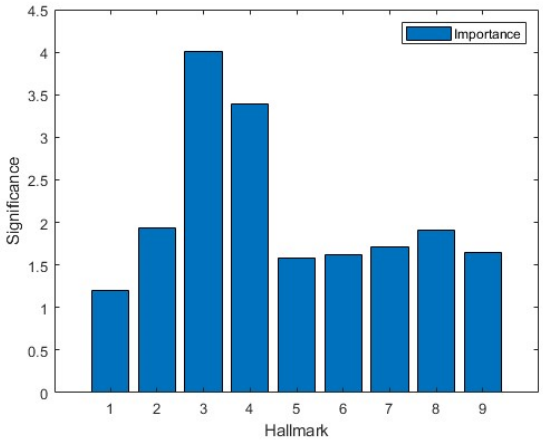

Figure 2. The feature importance parameters .
(Photo credit : Original)

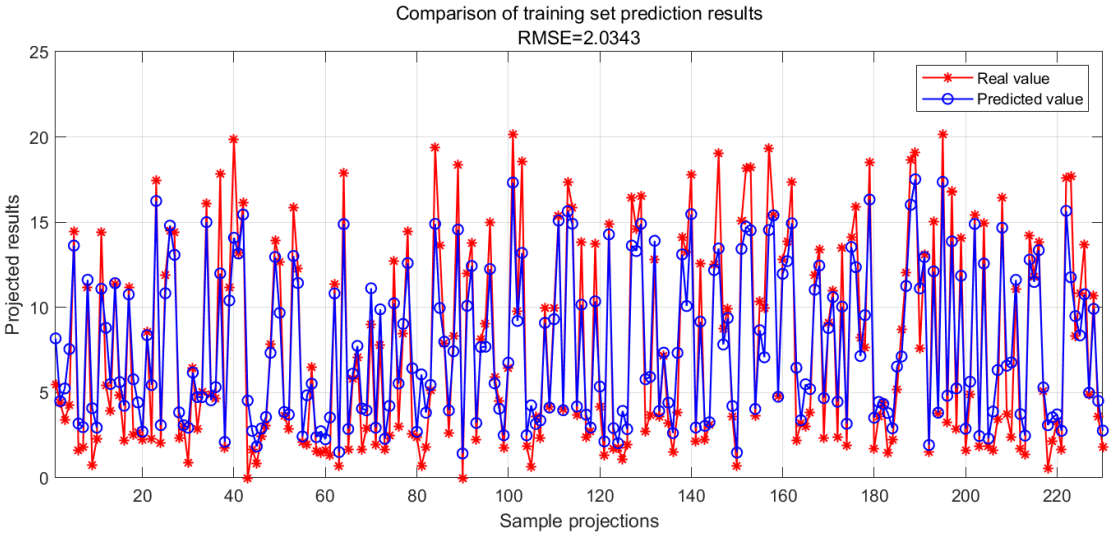

Figure 3. line graph of predicted-actual values for the training set.
(Photo credit : Original)

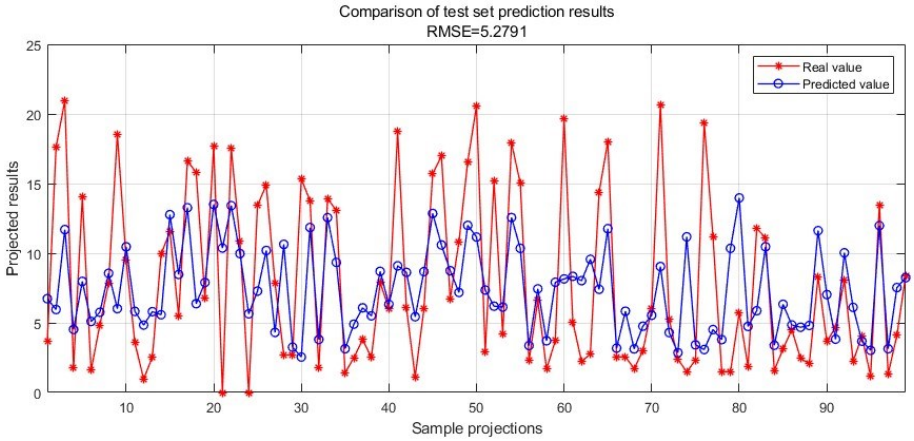

Figure 4. line graph of predicted-actual values for the test set.
(Photo credit : Original)

DOI: 10.54254/2755-2721/75/20240532

As can be seen from the line graph of predicted-actual values in the training set and the line graph of predicted-actual values in the test set, the prediction results of melanoma size in the training set are very close to the actual values, which indicates that the Random Forest regression algorithm based on particle swarm optimization is able to predict melanoma well. Meanwhile, the predicted values of melanoma in the test set are closer to the actual values. In this paper, R2, MAE and MBE are used to evaluate the prediction effect of the model, as shown in Table 3.

Table 3. Indicators for model evaluation.

	R2	MAE	MBE
Training set	0.88253	1.5682	-0.0056607
Test set	0.27783	4.058	-0.49925

As can be seen from the results, the MAE of the training set is 1.5682, and the MAE of the test set is 4.058, which indicates that the model is able to predict the size of melanoma very accurately, and in the future, the model can be further optimized to achieve a more accurate prediction of the size of melanoma.

6. Conclusion

As a malignant skin tumor, melanoma size prediction has been a subject of great interest in the medical field. A data regression prediction algorithm based on particle swarm optimization random forest provides a new idea for accurate prediction of melanoma size. The algorithm is effectively applied to the size prediction of melanoma on the skin surface and other parts of the skin by modeling the development of melanocytes.

In this study, the dataset is divided into training and test sets in a ratio of 7:3. By analyzing the prediction results of the training and test sets, it is found that in the training set, the predicted values of melanoma size are highly consistent with the actual values, while in the test set, the predicted values are also closer to the actual values. Specifically, the mean absolute error (MAE) is 1.5682 in the training set and 4.058 in the test set, which indicates that the random forest regression algorithm based on particle swarm optimization is able to accurately predict the size of melanoma and has a high generalization ability.

Combined with the experimental results, it can be seen that the algorithm not only performs well in fitting the real values on the training set, but also achieves satisfactory results on the unseen test data. This provides a solid foundation for further optimizing the model and improving the prediction accuracy in the future. By continuing to improve the model structure, increase the sample size, and optimize the parameters, we are confident that we will be able to predict the size of melanoma more accurately and provide stronger support for related clinical diagnosis and treatment.

In conclusion, the particle swarm optimization-based random forest algorithm presented in this paper brings new ideas and effective tools to the field of melanoma size prediction. With further in-depth research and technical improvement, it is believed that this method will become one of the indispensable and important tools in melanoma-related medical fields in the future.

References

[1] Aggarwal, R., & Goodell, J. W. (2010). Financial markets and the role of government in the Korean miracle. journal of Asian Economics, 21(2), 140-149.

[2] Wu, B. X., Zhan, G. S. (2018). A comparison of Hong Kong and Singapore financial centers. Journal of Economic Research, (31) 4, 86-87.

[3] Aizenman, J., Chinn, M. D., & Ito, H. (2010). The emerging global financial architecture: tracing and evaluating new patterns of the trilemma's configurations. journal of international money and Journal of International Money and Finance, 29(4), 615-641.

DOI: 10.54254/2755-2721/75/20240532

[4] Berger, A. A., & Huntington, H. G. (2005). The world's business cultures: and how to unlock them. Kogan Page Publishers.
[5] Chan, K., & Hameed, A. (2006). Stock price synchronicity and analyst coverage in emerging markets. journal of financial economics, 80(1), 115-147.
[6] Bob Chin (2016). International financial center competition: a comparison of Hong Kong and Singapore. International Economic Review, (42) 6, 67-78.
[7] Pan, Chunmei (2014). Comparative analysis of the financial center status of Hong Kong and Singapore. Research on Financial and Economic Issues, (28)2, 45-54.
[8] Huang, R. D., & Ito, T. (2012). Monetary policy and asset prices revisited. journal of financial economics, 106(3), 398-417.
[9] Chen, X. C., Lin, Y. H. (2012). A comparative study of Hong Kong and Singapore international financial centers. Finance China, (10), 39-45.