Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 3690077, 10 pages
https://doi.org/10.1155/2022/3690077

Research Article

Movie Box Office Prediction Based on IFOA-GRNN

Wei Lu, Xiaoqiao Zhang, and Xinchen Zhan

State Key Laboratory of Media Convergence Communication, School of Economics and Management, Communication University of China, Beijing 100024, China

Correspondence should be addressed to Wei Lu; luwei@cuc.edu.cn

Received 21 June 2022; Revised 8 July 2022; Accepted 16 July 2022; Published 21 August 2022

Academic Editor: Wen-Tsao Pan

Predicting movie box office has received extensive attention from academia and industry. At present, the main method of forecasting movie box office is subjective prediction, which is not widely accepted due to its accuracy and applicability. This study improves the fruit fly algorithm to optimize the generalized regression neural network (IFOA-GRNN) model to predict whether a movie can become a high-grossing movie. By using the actual box office data and performing virtual simulation calculations, the root means square error of the IFOA-GRNN model predicting the movie box office is 0.3412, and the classification accuracy is about 90%. By comparing this model with FOA-GRNN, KNN, GRNN, Random Forest, Naive Bayes, Ensembles for Boosting, Discriminant Analysis Classifier, and SVM, it is found that the prediction effect of the IFOA-GRNN model is significantly better than the above eight models. The contribution of this article is to propose a generalized regression neural network model based on an improved fruit fly optimization algorithm, which can greatly improve the accuracy of movie box office prediction.

1. Introduction

With the further advancement of China's modernization and the prosperity of pan-entertainment, movies have become the mainstay of China's cultural industry [1]. Regardless of age, occupation, and region, people now regard movies as an indispensable part of their spiritual life. At present, Chinese film market has become larger and larger [2, 3]. The total box office of the global film market in 2021 has reached 21.4 billion US dollars (Cinema and Home Entertainment Market Report 2021). Compared with 2020, which was affected by the pandemic, the global film market in 2021 has picked up significantly. Moreover, in 2021, China has become the most grossing country in the world, accounting for 34% of the world's box office (Analysis report by UK investment agency Gower Street (2021)), indicating the vigorous development of the Chinese film market in the postpandemic era.

Box office is an important indicator of the success of a movie. Usually, we evaluate a movie by its box office. For producers, making a high box office movie can accumulate a good reputation and pave the way for future creations. For investors, a high box office movie can bring a lot of benefits [4]. At the same time, an expected high-grossing movie can reduce the risk of investment, so they are more concerned about box office revenues than before in the current situation when the world is hit by a new coronavirus outbreak. For general movie audience, a high-grossing movie can enrich their spare time. Therefore, the box office level of a movie attracts more and more people's attention.

2. Literature Review

The influencing factors of movie box office and box office prediction are popular research fields in recent years. The box office of a movie may be affected by a combination of factors such as budget and film schedule. [5] Regarding the factors, Barry [6], who was the first to study movie box office prediction, used variables such as movie type, director, and actor influence, sequel or not when constructing a multiple regression movie box office prediction model. In the current study, Zhou and Han [7] believed that movie box office is significantly related to the number of reviews and user attention. Desai and Basuroy [8] found that for cultural products, such as books, music, and movies, the category of the product has a significant impact on its popularity. The

higher the star appeal and positive comments, the popularity of cultural products has a significant positive promotion effect. Zhao and Gao [9] divided films into cultural capital and social capital and found that the former starring and director's IP adaptation had a positive impact on the film box office, while the latter's reputation and policy factors had a positive impact on the film box office. Dhar et al. [10] used 26 years of domestic movie box office data and based on controlling for other factors affecting the movie box office, they concluded that series movies have higher box office compared with other movies without sequels. Kim et al. [11] divided word-of-mouth into expert comments and public comments to explore the impact of word-of-mouth on movie box office and found that positive expert comments and public comments had a positive effect on the box office, with expert comments having a higher value than public comments.

For the method of movie box office prediction, Lin and Liu [12] used the conventional multiple linear regression method, Dai and Zheng [13] used the OLS regression analysis method, and Xi [14] used machine learning algorithms such as xgboost to predict the movie box office. Song et al. [15] predicted movie box office based on random forest and found that the prediction effect was better than multiple linear regression. Lu and Xing [16] used the forward feedback neural network to predict the movie box office interval and used the regression method to predict the movie box office of the first week. Luo et al. [17] used the two-step system GMM to establish a movie box office prediction model, and the average absolute error percentage was about 9%. Atk [18] input social network service (SNS) data as a variable into a traditional machine learning model to predict movie box office, and the results show that the effect is significantly better than a single machine learning model.

For the above methods, first of all, the accuracy of the linear regression method is not high, and the movie box office and the influencing factors are not a simple linear mapping relationship. In addition, for nonlinear fitting methods, most of them are related to self-adjusting parameters, such as the penalty parameters and kernel parameters of the SVM classifier, the number of neurons in the neural network, the number of subtrees in the random forest, and the minimum sample leaf size. A large number will lead to a long time for algorithm parameter tuning, which will affect the real-time performance of prediction [19]. The generalized regression neural network model is suitable for solving nonlinear problems such as movie box office prediction. The model has the advantages of no training weights, fast learning speed, and single adjustment factor [20]. Moreover, the adjustment factor of the generalized regression neural network method has only one spreading parameter σ, and the structure is simple, and the nonlinear mapping ability is strong. It belongs to the method with fast learning speed and strong approximation ability, so it is suitable for the prediction of movie box office.

In this article, we proposed a generalized regression neural network optimized by the fruit fly optimization algorithm [21]. The dataset is divided into training set and test set, and the box office data are discretized into binary data by binning, namely, high box office and low box office. We analyzed to find out the many factors that affect the movie box office, and finally establish a variety of classification models. In addition, we compared the predicted results with the real values and found that common classifiers generally have problems such as insufficient prediction and poor stability. In a word, the current neural network prediction model is improved by using the fruit fly optimization algorithm, a branch of the emerging swarm algorithm.

This study proposes a generalized regression neural network model based on the improved fruit fly optimization algorithm (IFOA-GRNN), using the improved fruit fly optimization algorithm in the swarm algorithm to optimize the adjustment factor, and using the ten-fold cross-validation [22] method MATLAB calculation is carried out to effectively improve the prediction ability of the model. The fruit fly optimization algorithm has the advantages of simple method and fast convergence speed [23], which can quickly find the optimal value of the adjustment factor and improve the prediction accuracy [24]. In addition, compared with the traditional fruit fly optimization algorithm, the improved fruit fly optimization algorithm studied in this article has the advantages of being able to jump out of the local optimal solution. Through the comparison on the TMDB movie box office dataset, the proposed method achieves better classification performance. Compared with the existing traditional machine learning methods such as SVM and Random Forest, the accuracy, and precision are improved by about 20% on average.

3. The Fruit Fly Optimization Algorithm and Parameter Design

3.1. Symbol Description. 3.2. Traditional Fruit Fly Optimization Algorithm. Fruit fly optimization algorithm (FOA) is an emerging optimization algorithm [24, 25]. The algorithm is based on the foraging process of fruit flies. Fruit fly is a very sensitive creature with a very sensitive sense of smell, and it can even smell odors 40 kilometers away. Once the fruit fly smells the food, it can fly in its direction, and its foraging process is shown in Figure 1. The specific steps of the fruit fly optimization algorithm are as follows:

(1) Randomly initialize the position InitX_axis and InitY_axis of the fruit flies.

(2) Give the flies a random direction and distance to search for food.

$$\begin{cases} X_i = X\text{\_axis} + \text{Random Value}, \\ Y_i = Y\text{\_axis} + \text{Random Value}. \end{cases} \tag{1}$$

(3) Calculate the distance of the fruit fly from the origin D_i and the judgment value of taste concentration S_i.

$$\begin{aligned} S_i &= \frac{1}{D_i}, \\ D_i &= \left(X_i^2 + Y_i^2\right)^{1/2}, \\ \text{Smell}_i &= \text{Function}\left(S_i\right). \end{aligned} \tag{2}$$

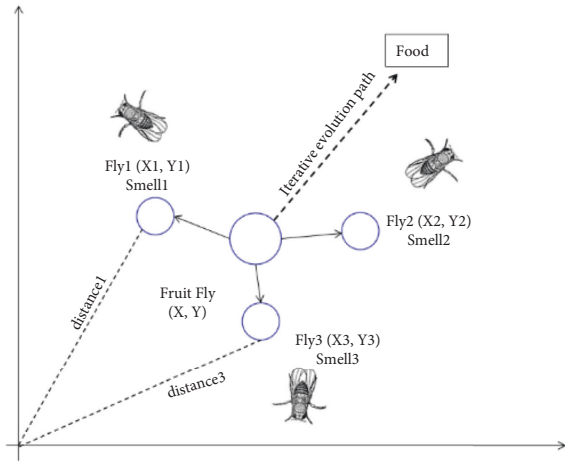

Figure 1: Fruit fly foraging pathway.

(4) Find the fruit fly with the highest concentration of taste in the whole fruit fly population and record its location.

$$[\text{best Smell bestindex}] = \max(\text{Smell}). \tag{3}$$

(5) The coordinates with the highest taste concentration are recorded and the flies swarm to it.

$$\text{Smell best} = \text{best Smell}. \tag{4}$$

(6) Enter the iterative optimization stage and repeat steps 2 to 5. If the smell concentration of the current position is higher than that of the previous iteration, execute the sixth step. Until the iteration ends, or the convergence conditions are met, the highest flavor concentration is finally obtained.

$$\begin{cases} X\_\text{axis} = X(\text{bestIndex}), \\ Y\_\text{axis} = Y(\text{bestIndex}). \end{cases} \tag{5}$$

4. IFOA-GRNN Model Design

4.1. GRNN Model. In 1991, Specht [26] proposed generalized regression neural network (GRNN), which is a special form of radial basis function neural network (RBF) [27]. GRNN is a feedforward neural network model based on nonlinear regression theory, which can handle nonlinear or linear regression problems well, and approximate functions by activating neurons. The generalized regression neural network consists of four layers, namely, the input layer, the pattern layer, the summation layer, and the output layer. Its structure is shown in Figure 2. The generalized regression neural network does not need to be trained, and its network link weight value is determined by the output and input of the training sample, and there is no need to estimate and guess the number of hidden layers and hidden units in the network [28]. It is derived from RBF and therefore has only one free parameter, the RBF smoothing parameter. It can be seen that the purpose of data mining is to find the optimal smoothing parameter σ and the number of network neurons.

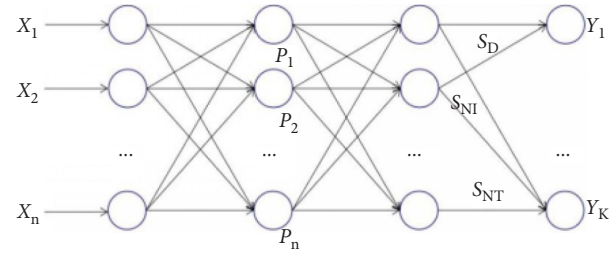

Figure 2: Generalized regression neural network structure diagram.

The theoretical basis of GRNN is nonlinear regression analysis, which has strong nonlinear mapping ability, and because it does not need to train weights, its learning speed is also very fast [29].

4.1.1. Input Layer. The number of neurons in the input layer is equal to the dimension of the input vector in the sample, and the input layer passes the vector directly to the pattern layer.

4.1.2. Pattern Layer. The number of neurons in the pattern layer is equal to the number of samples, and the transfer function of the pattern layer is

$$P_i = e^{\left[-\left((X-X_i)^T(X-X_i)/2\sigma^2\right)\right]}. \tag{6}$$

4.1.3. Summation Layer. The summation layer has two types of neurons that use different formulas for summation.

The first type of summation is the arithmetic summation of all pattern layer inputs.

$$S_D = \sum_{i=1}^{n} P_i. \tag{7}$$

The second summation method is to perform a weighted summation of all pattern layer inputs, and the connection weight between the pattern layer and the summation layer is the j element in the i output sample.

$$S_{Nj} = \sum_{i=1}^{n} y_{ij} P_i \quad j = 1, 2, 3, \ldots, k. \tag{8}$$

4.1.4. Output Layer. The number of neurons in the output layer is equal to the dimension of the vector output in the sample, and each neuron divides the output of the summation layer to obtain the final output.

$$y_j = \frac{S_{Nj}}{S_D}. \tag{9}$$

It can be seen from the above formula that the generalized regression neural network does not need to be trained, and the only parameter that GRNN needs to determine is σ, so how to choose a suitable σ is very important for our

prediction work. If the value of σ is too large, the value of is the output in formula (9) approximately the average value of the sample data, and if the value of σ is too small, y_j will be too close to the sample data, and the generalization ability of the model will be greatly reduced. Therefore, this study chooses the IFOA [30–32] to find the optimal spreading parameter σ.

4.2. GRNN Model Optimized by IFOA. The traditional fruit fly optimization algorithm FOA has problems such as easy to fall into local optimum [33], too many iterations, and too long search time. For the search step size of FOA that is relatively fixed, if it is too large, the optimal solution may be missed, and the local search ability is relatively weak. If it is too small, the global search ability is relatively weak, and so the optimal solution may be missed as well. At the same time, a positive search step size in FOA means that the fruit fly may only search in one direction, and negative values should not be ignored.

Based on the above problems, to solve the spreading parameter σ in the generalized regression neural network, this study adopts an improved fruit fly optimization generalized regression neural network (IFOA-GRNN) method to achieve our purpose. The main idea of the improvement is to give a larger step size in the first half of the iteration, while adjust to a smaller step size in the second half of the iteration for precise search and use the sign function (sgn) as the step size on the basic search step size. Adding plus and minus signs can increase the diversity of foraging processes in fruit fly.

IFOA to optimize GRNN is shown in Figure 3.

Step 1. Randomly initialize the position of the fruit fly InitX_axis, InitY_axis.

$$\begin{cases} X\_\text{axis} = \text{rand} \\ Y\_\text{axis} = \text{rand} \end{cases}$$

$$L_0 = \text{sign}\,(2 * \text{rand} - 1)$$

$$L = \begin{cases} L_0 * \left(1 + \dfrac{\text{gen}}{\max \text{gen}}\right), & if\ 0 < \dfrac{\text{gen}}{\max \text{gen}} <= 0.5, \\ L_0 * \left(1 - \dfrac{\text{gen} - 1}{\max \text{gen}}\right), & if\ 0.5 < \dfrac{\text{gen}}{\max \text{gen}} <= 1, \end{cases} \tag{10}$$

$$\begin{cases} X_i = X\_\text{axis} + L, \\ Y_i = Y\_\text{axis} + L. \end{cases}$$

Step 2. Calculate the distance from the origin of the fruit fly D_i and the judgment value S_i of taste concentration, which is the spreading parameter σ in the GRNN network.

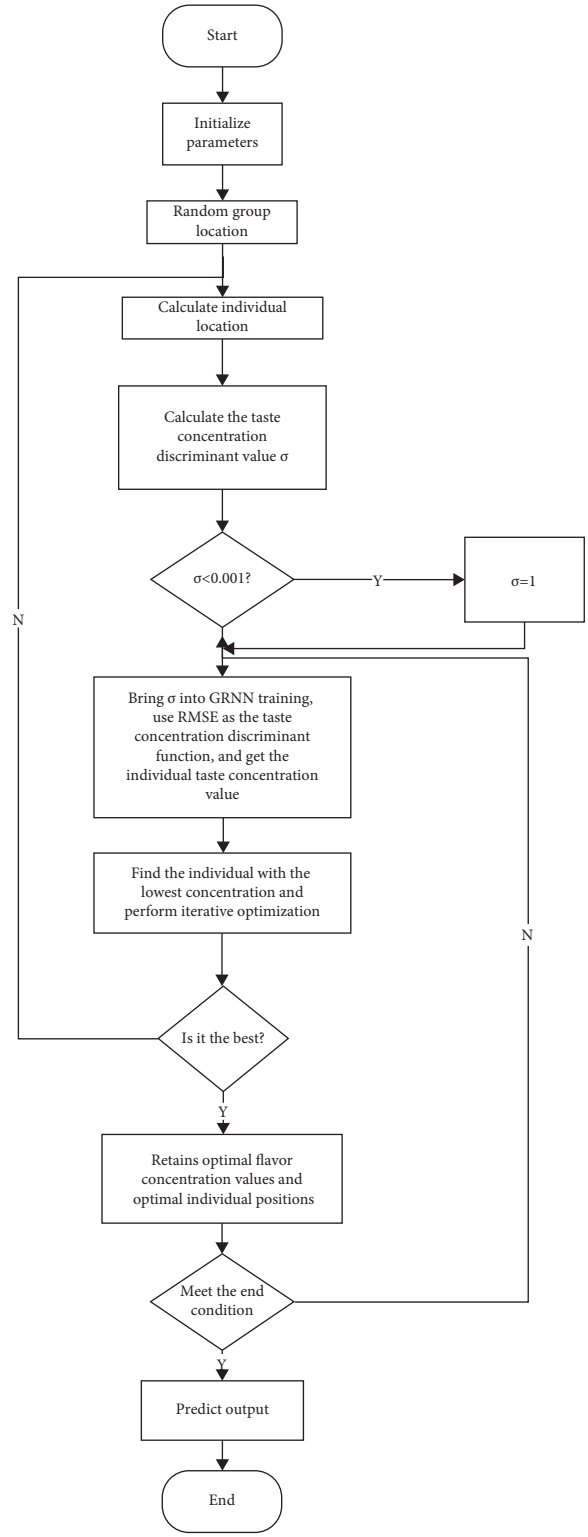

Figure 3: IFOA-GRNN algorithm flow chart.

$$D_\text{i} = \left(X_\text{i}^2 + Y_\text{i}^2\right)^{1/2},$$

$$S_i = \frac{1}{D_i}. \tag{11}$$

Step 3. Substitute the taste concentration judgment value S_i of the fruit fly at this position into the taste concentration judgment function Smell_i to obtain the taste concentration of the fruit fly at this position. The taste concentration determination function here is the error function, which is used to calculate the RMSE. The lower the RMSE value, the better the model fitting effect.

$$\begin{aligned} \text{Smell}_i &= \text{Function}(S_i), \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{k}^{n} y(k) - y}, \\ [\text{bestSmell bestindex}] &= \min(\text{Smell}). \end{aligned} \tag{12}$$

Step 4. Record the coordinates with the lowest taste concentration and the flies swarm to it.

$$\begin{aligned} \text{Smell} &= \text{bestSmell}, \\ \begin{cases} X\_\text{axis} = X(\text{bestIndex}), \\ Y\_\text{axis} = Y(\text{bestIndex}). \end{cases} \end{aligned} \tag{13}$$

5. Empirical Research and Data Analysis

5.1. Data Source. Movie box office is affected by many factors. Through the significance test, this study selects seven main influencing factors: budget, production country, production company, genres, runtime, homepage, and popularity, to predict movie box office. The data source of this article is the Movie Database (TMDB) dataset provided by Kaggle (https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata), which includes the information of nearly 5,000 movies. There are 23 variables, such as movie_id (TMDB movie identification number), title (movie name), cast (actor list), director (director), genres (style list/movie type), and homepage of a movie (Movie front page).

5.2. Data Processing. The original data contains twenty-three attributes such as revenue, budget, keywords, and so on. After a preliminary screening to remove meaningless indicators (such as poster URLs) and data that are difficult to quantify (such as movie descriptions), budget, genres, popularity, production company, production country, runtime, homepage, and movie language are eight attributes. As the purpose of this study is to predict whether a movie is a high box office movie or a low box office movie, it is necessary to use the binning operation to discretize the box office indicators. To avoid the influence of outliers, this article chooses to use equal-frequency binning. The same number of samples is in the "low box office" set and the "high box office" set. We use 1 for "high box office" and 0 for "low box office."

In addition, genres, production companies, and production countries are all unstructured data. These indicators are very difficult to quantify and analyze, and there is no unified judgment standard. Therefore, this study takes the weighted sum of the average rating and the average number of reviews corresponding to previous films of this type as the rating of the genres:

$$\text{genres} = \frac{\text{std}(\text{review\_score}) + \text{std}(\text{review\_num})}{2}. \tag{14}$$

The following calculation method is the same for the production companies and production countries:

$$\begin{aligned} \text{companies} &= \frac{\text{std}(\text{review\_score}) + \text{std}(\text{review\_num})}{2}, \\ \text{countries} &= \frac{\text{std}(\text{review\_score}) + \text{std}(\text{review\_num})}{2}. \end{aligned} \tag{15}$$

Then, delete the row where there are abnormal values (such as zero or null) in the first eight attributes and normalize the sample data to exclude the influence of dimension, so that the data are between 0 and 1, and finally get the data we need.

5.3. Indicator Selection. After data processing, there are initially eight indicators that may be related to movie box office. In order to obtain the final required influencing factors, this study uses MATLAB to conduct the significance test method to test the significance of the following eight indicators after data processing: budget, genres, popularity, production company, production country, runtime, homepage, and movie language. The results shown in Table 1 are obtained.

It can be seen from Table 2 that the seven variables of budget, genres, popularity, production company, production country, runtime, and homepage passed the significance test, indicating that they are all related to the movie box office, and the variable of movie language did not pass the significance test. After screening out the amount, we finally selected the first seven factors as the seven indicators for the movie box office prediction in this study.

5.3.1. Budget. Although there are low-cost, high box office movies such as "Chainsaw" and "Crazy Stone," various phenomena show that the scale of investment in a film is proportional to its box office. High investment can guarantee the production intensity of all aspects of the film, thereby creating the possibility of high box office. Coupled with the popularity of the Internet and piracy, audiences will think that there is no big difference between watching low-cost movies on computer TV and watching movies in theaters, which makes it even more difficult for low-cost movies to get the box office.

5.3.2. Genres [34]. An appealing script will undoubtedly lead to a high box office. For example, the classic thriller "Chainsaw" series, each of which is a low-budget production, has gained wide popularity because of its fascinating and logical script settings, and the series of films have been released one after another. With the development of film

Table 1: Significance test table of box office influencing factors.

Factor	Factor number	T-test significance level
Budget	X1	0.000
Genres	X2	0.000
Popularity	X3	0.000
Production company	X4	0.000
Production country	X5	0.001
Runtime	X6	0.000
Homepage	X7	0.000
Movie language	X8	0.614

Table 2: The definitions of symbols.

Symbol	Definition
X	Fruit fly abscissa
Y	Fruit fly ordinate
X_axis	Fruit fly starting abscissa
Y_axis	Fruit fly starting ordinate
D	Fruit fly distance from origin
S	Taste concentration judgment value
Smell	Taste concentration
bestSmell	Maximum taste concentration
bestIndex	Fruit fly location with greatest flavor concentration
L	Fruit fly distance per random search
Σ	Spreading parameters
RMSE	Root mean square error

industrialization and marketization, the type of film has gradually become apparent, and the relationship between type and box office has also deepened. Today, classic film genres such as action films, disaster films, children's films, horror films, romantic films, and so on have been formed. Each type of film corresponds to a different audience group, and the spending power of each audience group is different. Therefore, the film genre has an impact on the box office of the film. This article uses the word cloud to show the distribution of movie genre (Figure 4).

5.3.3. Production Companies. A good production company is often one of the guarantees of a high-grossing film, such as Universal Pictures, Columbia Pictures, Warner Bros. Entertainment, The Walt Disney Company, and so on. Famous film production companies are often able to use their own resources to choose a better release date, have a stronger cast and director lineup, and a more mature production model, all of which can contribute to the success of a film.

5.3.4. Production Countries. The origin of the movie also affects the audience of the movie to a certain extent, as the origin of the movie invisibly gives people a subjective awareness and emotion about the national culture [35]. We can also find that many high box office movies are produced in the United States, China, and other countries.

5.3.5. Popularity. Popularity refers to the relative number of clicks of the movie in the Movie Database, and we can also understand it as its relative number of searches. The more times a movie is searched before it is released, the higher its attention [4], and this search has a certain probability that

Figure 4: Movie genre word cloud.

the audience actually goes to the cinema to watch the movie, which in turn affects the box office. Therefore, popularity is one of the important factors.

5.3.6. Runtime. From a psychological level, runtime is considered to be a key factor that can affect a movie's box office. Generally speaking, if a movie is too long, the audience may feel tired, and if a movie is too short, it may bring a perfunctory feeling to the audience's psychology, which will affect the reputation of the movie, and then affect the box office of the movie.

5.3.7. Homepage. The homepage variable refers to whether a movie has its own homepage before it is released. Many excellent movie producers will make a promotional homepage for the movie before the movie is released, which will contain starring information, trailers, and other content. The promotional cost of the movie has doubled in the past few years, which reflects the growing demand of audience for movie information. It also shows that the influence of publicity and promotion on the audience increases, and then it can be inferred that it has a positive impact on the movie box office.

5.4. Evaluation Indicators. In this study, the accuracy rate, precision rate, recall rate, and F1 value are used to evaluate the prediction effect of the model, as shown in Table 3.

Accuracy: As the name implies, it is the proportion of all correct predictions in the total.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{16}$$

Precision: That is the proportion of those who are correctly predicted to be high box office accounts for the total predicted high box office.

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{17}$$

Table 3: Confusion matrix.

Actual value	Predictive value	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Recall: It is the proportion of the total box office that is correctly predicted to be a high box office is actually a high box office.

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{18}$$

F1 value: Consider both precision and recall.

$$\frac{2}{F_1} = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}},$$
$$F_1 = \frac{2TP}{2TP + FP + FN}. \tag{19}$$

5.5. Experiment. After the TMDB_5000 dataset is screened and preprocessed, there are 3169 movie data remaining. Among them, 70% of the data is used as the training set, and 30% of the data is used as the validation set, that is, 2220 movie data are trained, and the model generalization ability is verified on 949 movie data. For the training set, this study divides it into 10 parts, 9 as the training set and 1 as the test set, followed by using 10 cross-validation methods to calculate the mean of the model evaluation indicators. We choose seven evaluation indicators of budget, genres, popularity, production company, production country, runtime, and homepage as influencing factors to build a model.

The network structure of GRNN in this model is well determined. It is a four-layer neural network. The number of neurons in the input layer is the dimension of the input vector, the number of neurons in the pattern layer is equal to the number of samples, and the number of neurons in the output layer is the dimension of the output vector. To make the IFOA-GRNN prediction model optimal, the population number and the maximum number of iterations of fruit flies were continuously adjusted in the experiment [36], and the optimal spreading parameter values under different parameters were obtained, as shown in Table 4.

Through experiments, it is found that when the fixed population size is 10 and the number of iterations is 100, the spreading parameter σ reaches the minimum value of 0.053. Although the spreading parameter can reach the optimal value of 0.0530 by increasing the population size or the number of iterations, it will increase the time for iterative optimization. At the same time, this article also found that changing different population sizes will lead to different RMSE results during convergence, but the difference is small, and IFOA-GRNN is not sensitive to the initial parameter selection. Therefore, this article chooses 10 as the population size and 100 as the maximum number of iterations.

Table 4: Optimal spreading parameters under different parameters.

Population size	Number of iterations	RMSE	Spreading parameter σ
1	50	0.3458	0.0531
5	50	0.3458	0.0531
5	100	0.3458	0.0531
10	100	0.3458	0.0530
20	100	0.3458	0.0530
30	100	0.3458	0.0530
30	150	0.3458	0.0530

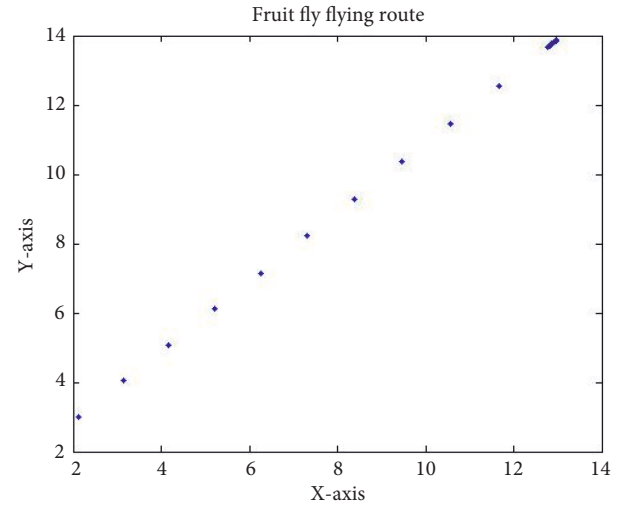

Figure 5: IFOA-GRNN fruit fly search path.

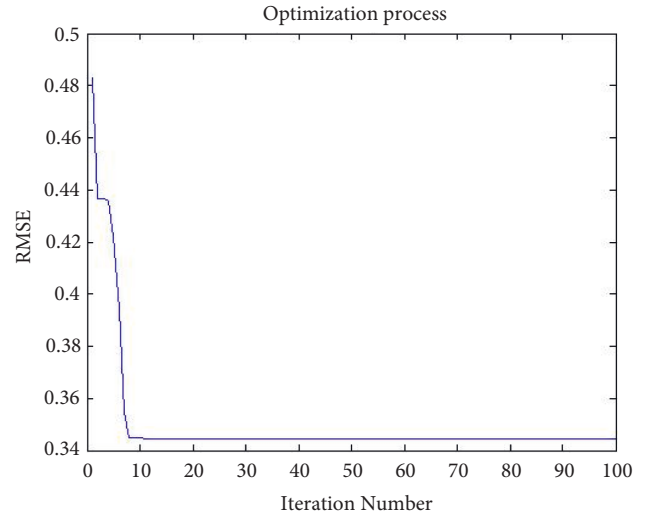

Figure 6: IFOA-GRNN convergence graph.

After the simulation, the search path of the fruit fly is shown in Figure 5, and the convergence of RMSE is shown in Figure 6. The IFOA-GRNN model finally converged in the ninth iteration after 100 iterations, with a minimum RMSE of 0.3458 and a spreading parameter σ of 0.0530.

For FOA-GRNN algorithm, the iterations converge to 9 generations. The optimal spreading parameter σ obtained is

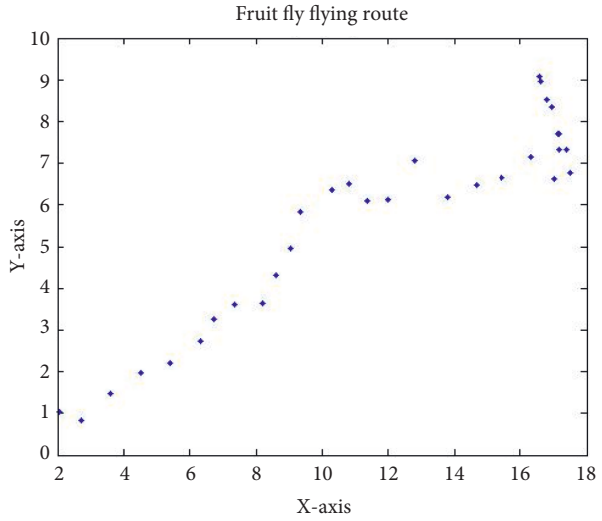

Figure 7: FOA-GRNN fruit fly search path.

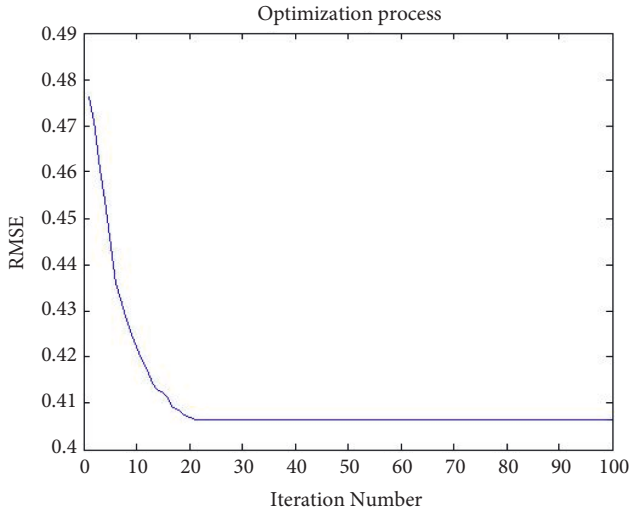

Figure 8: FOA-GRNN convergence graph.

Table 5: Prediction effect table of different classifiers.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 value
IFOA-GRNN	90.46	91.72	90.26	0.9098
FOA-GRNN	84.33	85.76	84.11	0.8492
GRNN	70.91	73.65	70.50	0.7204
KNN	73.55	73.91	73.66	0.7379
Naive Bayes	73.02	75.31	72.56	0.7389
Random forest	80.18	80.28	80.15	0.8022
Ensembles for boosting	78.61	78.83	78.69	0.7876
Discriminant analysis classifier	75.55	77.84	75.27	0.7653
SVM	79.18	79.35	79.08	0.7921

0.0 623, and the minimum value of RMSE is 0.4064. The optimal path of fruit fly is shown in Figure 7. The RMSE

convergence is shown in Figure 8.

Compared with the traditional FOA-GRNN model, the IFOA-GRNN model proposed in this article achieves the convergence faster, but also skips the local optimum, searches for the optimal spreading parameters, and has better prediction performance.

To comprehensively evaluate the prediction ability of IFOA-GRNN, this study chooses to use traditional classifiers such as KNN, GRNN, Naive Bayes, Random Forest, Ensembles for Boosting, Discriminant Analysis Classifier, and SVM to predict the dataset in this study, and it is consistent with the core of this study. The methods are compared and the results in Table 5 are obtained.

6. Conclusions

In this study, we used the fruit fly optimization algorithm to optimize the spreading parameter σ in the generalized regression neural network, which has achieved good results in the prediction and analysis of popular movies. Taking the TMDB_5000 dataset as an empirical research, this study compares the prediction results of the IFOA-GRNN model with FOA-GRNN, GRNN, KNN, and other models, and draws the following conclusions:

(1) This article conducts a significant T test on the sample before modeling, and finds that seven factors of budget, genres, popularity, production company, production country, runtime, and homepage significantly affect the box office of a movie.

(2) This article selects different fruit fly population numbers for experiments. It is found that when the population number changes, the results obtained by MATLAB simulation have little difference, indicating that the IFOA-GRNN model has strong stability and is not sensitive to the selection of IFOA initial parameters.

(3) In this article, IFOA is selected to automatically find the optimal spreading parameter σ of GRNN, which reduces the interference of artificial selection on prediction. At the same time, by comparing with the FOA-GRNN model, it is found that the IFOA-GRNN model can jump out of the local optimum and find the global optimum solution more quickly. Compared with classification models such as SVM and KNN, it is found that the accuracy, precision, recall, and F1 value of IFOA-GRNN are significantly higher than other classifiers, indicating that the model can significantly improve the accuracy of movie box office prediction.

Because IFOA-GRNN model combined with generalized regression neural network has good stability, accuracy, and computing speed, it can be applied to the research of prediction problems in machine learning, cognitive science, and other fields.

Based on the work of this study, the feasible improvement directions in the future include the following:

(1) This article selects the seven factors affecting the movie box office—budget, genres, popularity, production company, production country, runtime, and homepage. In fact, movie box office should be influenced by many factors, such as the cast and director lineup. Due to the difficulty of data collection, more factors are not considered in this article. In the future, more factors related to movie box office can be added to the model through further research.

(2) When choosing the optimal population size and the maximum number of iterations in the IFOA-GRNN algorithm, the empirical method is used to select some conventional combinations because the experimental computing power is limited, and there is no way to traverse to find a better combination of population size and maximum number of iterations that may exist. In future research, if there is a better experimental environment, we can consider solving the optimal combination of population size and maximum number of iterations. [37].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

This paper is one of the stage achievements of the State Key Laboratory of Media Convergence Communication and the research and cultivation project of Communication University of China, which is based on the big data film box office research.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the special funds of the basic scientific research service of the Central University in Communication University of China.

References

[1] S. Rao, "The power of infection: cultural soft power and film," *Contemporary Film*, no. 02, pp. 15–21, 2008.

[2] Z. Wang and M. Xu, "Analysis of influencing factors of movie box office——research based on logit model," *Exploration of Economic Problems*, no. 11, pp. 96–102, 2013.

[3] X. Hu, Li Bo, and Z. Wu, "Analysis of influencing factors of movie box office," *Journal of Communication University of China (Natural Science Edition)*, vol. 20, no. 01, pp. 62–67+39, 2013.

[4] M. T. Lash and K. Zhao, "Early predictions of movie success: the who, what, and when of profitability," *Computer Science*, vol. 33, no. 3, 2015.

[5] S. Lee and J. Y. Choeh, "The interactive impact of online word-of-mouth and review helpfulness on box office revenue," *Management Decision*, vol. 56, no. 4, pp. 849–866, 2018.

[6] B. R. Litman and L. S. Kohl, "Predicting financial success of motion pictures: the '80s experience," *The Journal of Media Economics*, vol. 2, no. 2, pp. 35–50, 1989.

[7] M. Zhou and D. Han, "A movie box office prediction model based on social media user comments and attention," *Microcomputer Applications*, vol. 33, no. 18, pp. 73–75, 2014.

[8] K. K. Desai and S. Basuroy, "Interactive influence of genre familiarity, star power, and critics' reviews in the cultural goods industry: the case of motion pictures," *Psychology and Marketing*, vol. 22, no. 3, pp. 203–223, 2005.

[9] X. Zhao and F. Gao, "A study on the influencing factors of China's main theme movie box office," *Film Literature*, no. 20, pp. 3–7, 2020.

[10] T. Dhar, G. Sun, and C. B. Weinberg, "The long-term box office performance of sequel movies," *Marketing Letters*, vol. 23, no. 1, pp. 13–29, 2012.

[11] S. Ho Kim, N. Park, and S. H. Park, "Exploring the effects of online word of mouth and expert reviews on theatrical Movies'Box office success," *The Journal of Media Economics*, vol. 26, no. 2, pp. 98–114, 2013.

[12] Q. Lin and J. Liu, "Box office prediction of my country's commercial films based on multiple linear regression," *Science and Technology and Economics Tribune*, vol. 28, no. 22, pp. 7–9, 2020.

[13] J. Dai and Y. Zheng, "A continuity study of the box office performance of movies," *Modern communication (Journal of Communication University of China)*, vol. 40, no. 08, pp. 124–129, 2018.

[14] Y. Xi, "Prediction of movie box office based on machine learning algorithm," *Electronic Production*, no. 04, pp. 51-52+55, 2021.

[15] Y. Song, J. Zhu, Q. Yang, Z. Fu, and Ke Xu, "The first-week box office prediction analysis of domestic films based on random forest regression," *Science Journal of Normal University*, vol. 41, no. 01, pp. 21–26, 2021.

[16] W. Lu and R. Xing, "Research on movie box office prediction model with conjoint analysis," *International Journal of Information Systems and Supply Chain Management*, vol. 12, no. 3, pp. 72–84, 2019.

[17] X. Luo, J. Qi, and C. Tian, "Research on the box office prediction model after the movie premiere," *Statistics and Information Forum*, vol. 31, no. 11, pp. 94–102, 2016.

[18] B. J. H. Atk, "Box office forecasting using machine learning algorithms based on SNS data," *International Journal of Forecasting*, vol. 31, no. 2, pp. 364–390, 2015.

[19] Pi Jun, S. Ma, Q. Zhang, L. Wang, and D. Cui, "GRNN aero-engine exhaust temperature prediction model based on improved fruit fly algorithm optimization," *Aerodynamics*, vol. 34, no. 01, pp. 8–17, 2019.

[20] Y. Tian, B. Zhang, D. Liu et al., "Prediction of state trend of GRNN hydropower unit based on fruit fly optimization algorithm," *Hydropower and Energy Science*, vol. 30, no. 12, pp. 127–129+89, 2012.

[21] W. Pan, "Application of fruit fly optimization algorithm to optimize generalized regression neural network for business performance evaluation," *Journal of Taiyuan University of Technology*, vol. 29, no. 04, pp. 1–5, 2011.

[22] Y. Feng, N. Cui, D. Gong, Q. Zhang, and L. Zhao, "Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling," *Agricultural Water Management*, vol. 193, pp. 163–173, 2017.

[23] W. T. . Pan, "A new Fruit Fly Optimization Algorithm: t," *Knowledge-Based Systems*, vol. 26, no. 2, pp. 69–74, 2012.

[24] J. Pan and T. Pan, *A New Evolutionary Computation Approach: Fruit Fly Optimization Algorithm*, 2011.

[25] W. T. Pan, C. E. Huang, and C. L. Chiu, "Study on the performance evaluation of online teaching using the quantile regression analysis and artificial neural network," *The Journal of Supercomputing*, vol. 72, no. 3, pp. 1–15, 2016.

[26] D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.

[27] J. Hou, X. Lu, K. Zhang et al., "Parameters identification of rubber-like hyperelastic material based on general regression neural network," *Materials*, vol. 15, no. 11, p. 3776, 2022 May 25.

[28] M. Huang and J. Liu, "Comparison of the application of GRNN and statistical models in monitoring seawall seepage pressure," *Hydropower and Energy Science*, vol. 27, no. 03, pp. 134–136, 2009.

[29] C. Zhao, K. Liu, and D. Li, "Freight volume forecast based on generalized regression neural network," *Journal of Railways*, no. 01, pp. 12–15, 2004.

[30] Z. Zhang and C. Liu, "Short-term load forecasting of power system based on IFOA-GRNN," *Energy Research and Information*, vol. 36, no. 03, pp. 162–166+178, 2020.

[31] D. Li, H. Yin, and B. Zheng, "Annual power load forecast based on MFOA-GRNN model," *Power Grid Technology*, vol. 42, no. 02, pp. 585–590, 2018.

[32] X. Zhu, "Research on short-term power load forecasting method based on IFOA-GRNN," *Power System Protection and Control*, vol. 48, no. 09, pp. 121–127, 2020.

[33] Y. Wang, N. Nie, M. Wang, and Z. Li, "Modified fruit fly algorithm to optimize GRNN network for tailings pond safety prediction," *Computer Engineering*, vol. 41, no. 04, pp. 267–272, 2015.

[34] Z. Han, B. Yuan, C. Yan, N. Zhao, and D. Duan, "An effective early movie box office prediction model based on GBRT," *Computer Application Research*, vol. 35, no. 02, pp. 410–416, 2018.

[35] S. Delre, "Simulating the cinema market: how cross-cultural differences in social influence explain box office distributions (version 4)," *Comses Computational Model Library*, vol. 4, 2010.

[36] X. Zhang, "Application of generalized regression neural network based on Drosophila algorithm in financial early warning," *Finance and Accounting Monthly*, vol. 37, no. 30, pp. 91–94, 2016.

[37] C. Liu, L. Zhang, J. Wang, and L. Li, "Prediction of water content of crude oil based on FOA-GRNN oil well measurement," *Computer Simulation*, vol. 29, no. 11, pp. 243–246+259, 2012.