An Interval Reliability Demand Prediction Method Combined with XGBoost and D-S Evidence Theory in Film Preparation Period

Zhongjun Tang and Lang Ni*

School of Economics and Management, Beijing University of Technology, Beijing 100124, China
Email: 13132587357@163.com

Abstract. The lack of historical sales data and word-of-mouth information in the film preparation period, the few available variables and the uncertainty in the prediction process lead to the difficulty in predicting the total box office demand of films. To solve this problem, this paper constructed and verified the prediction method of interval reliability demand in the film preparation period, which combined XGBoost algorithm and D-S evidence theory. Firstly, the total box office interval was effectively divided according to the sample data of the training set, and XGBoost was used to complete the calculation of the reliability function value of the evidence variables. Then, the D-S evidence theory was used for information fusion to obtain the results of box office interval reliability fusion. Finally, the box office attribution was judged by the interval reliability, so as to realize the interval reliability demand prediction in the preparatory period. The validity of the proposed method was verified by selecting the data of Chinese films from 2017 to 2019, and it was compared with the classical predictive classification algorithm. The results showed that the method has higher prediction accuracy and better generalization ability.

Keywords. Demand forecasting; XGBoost; D-S evidence theory; information fusion.

1. Introduction

With the rapid development of economy and the gradual improvement of people's living standard, the box office of domestic films is generally on the rise [1]. However, relevant data show that Chinese film box office polarization phenomenon is more and more obvious. Many of the films that were released were unable to recoup their costs, bringing huge economic losses to investors. In the whole life cycle of a film, investment and production decisions occur during the preparatory period [2]. However, in the preparation period, only the quality information of the product itself can be obtained [3], and there is a lack of marketing data and film reputation, as well as less available predictive information and uncertainty in the prediction process. These limitations bring great difficulties to the realization of the box office prediction in the preparation period.

In recent years, the rapid development of big data, data mining and information fusion provides a new way to solve such problems. Considering that the film production has not been completed in the preparatory period, the future earnings only need to be roughly evaluated to judge whether there will be a loss of investment, so the interval forecast is more consistent with the forecast requirements of this stage than the point value forecast. Focusing on this, relevant scholars began to construct different methods to realize the interval prediction in the preparatory period. However, in their studies, the methodological shortcomings are more obvious. For example, Ghiassi [4] and Zheng [5] proposed the
A box office prediction model based on neural network, which achieved a good prediction effect. However, the prediction interval reliability is not given in the results, and the results lack credibility. Ahmed [6] and Parimi [7] used classical machine learning algorithms such as SVM and GBDT to establish the classification prediction model, and the method showed good effectiveness. However, it failed to reflect the importance of features in the research process and failed to solve the uncertainty problem in the forecasting process. Tang [8] proposed to combine rough set with evidence theory, through the information fusion, the problem of uncertainty was solved. However, the data processing process based on rough set is relatively complex, and once the calculation error occurs in a certain step, it will bring some errors to the later fusion.

In order to solve the shortage of existing research and realize the prediction of interval demand in the preparation period, this paper proposed a prediction method of interval demand in the preparation period combining XGBoost and D-S evidence theory. Firstly, data of each variable is obtained through data mining and quantified. Secondly, the XGBoost algorithm is not affected by the data size and is not easy to produce the ability of overfitting. The quantitative data is predicted to obtain the prediction results of the single variable interval, and the reliability calculation is completed. Then, D-S evidence theory is introduced, each variable is regarded as independent evidence, and the fusion formula of evidence theory is used to fuse the reliability of each evidence, and the final fusion result is obtained to solve the uncertainty problems in the forecasting process. Finally, an example is given to illustrate the effectiveness of the method.

2. Method Building

2.1. Overall Process of Method

The overall process of the prediction method of interval reliability demand in the film preparation period combined XGBoost and D-S evidence theory was shown in figure 1, which mainly included three stages: data processing, reliability calculation and information fusion.

![Figure 1. Overall process of method.](image)

2.2. Data Processing

In the data processing stage, the difficulty of data crawling of various movie websites and the experimental needs are comprehensively weighed. The factors that affect the quality of film production, such as director, screenwriter, film type, first leading actor and second leading actor, were selected as variables, and the relevant data were crawled through the movie website, which were preprocessed and quantified.
2.3. Reliability Calculation
As a member of the ensemble algorithm, XGBoost has a good application prospect in the field of demand prediction due to its powerful classification learning and feature learning capabilities [9]. In this paper, the algorithm was used to classify and predict a single variable, and the interval prediction result was taken as the reliability function value of the variable. The algorithm flow is as follows.

(1) First initialize a function \( f_t(x) \), as shown in equation (1).

\[
f_t(x) = \omega_q(x) \quad \omega \in \mathbb{R}^T, \quad q: \mathbb{R}^d \rightarrow \{1, 2, \ldots, T\}
\]  

where, \( q(x) \) refers to the tree structure, represents the sample \( x \) is on the leaf node, \( \omega \) represents the weight sum of the leaf node of the decision tree, and is also the predicted value of the decision tree. The penalty function \( \Omega(f_t) \) is initialized as shown in equation (2).

\[
\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2
\]

where, \( T \) represents the number of leaf nodes, \( \gamma \) and \( \lambda \) represent the penalty coefficient.

(2) The square error function is set as the loss function, as shown in equation (3).

\[
L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2
\]

here, \( y_i \) is the true value and \( \hat{y}_i \) is the predicted value.

(3) Weak regression tree generation based on CART. First, the objective function is defined, as shown in equation (4).

\[
Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)
\]

Since the XGBoost prediction model is an addition model and the prediction result is the combination result of the current model and the new tree, the predicted value of the \( t \)-th iteration can be obtained by minimizing the following objective function, as shown in equation (5).

\[
Obj^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{const}
\]

The second-order Taylor expansion of the error function was substituted into the penalty function \( \Omega(f_t) \), which can be optimized as shown in equation (6).

\[
Obj^{(t)} = \sum_{j=1}^{T} \left[ g_j \omega_j + \frac{1}{2} (H_j + \lambda) \omega_j^2 \right] + \gamma T
\]

where, \( g_j = \sum_{i \in I_j} g_t \), \( H_j = \sum_{i \in I_j} h_i \) ( \( i \) is the sample set on the parent node) respectively represented as the sum of the first-order and second-order gradients in each leaf node. \( g_t = \partial \gamma^{(t-1)} L(y_i, \hat{y}^{(t-1)}) \), \( h_t = \partial^2 \gamma^{(t-1)} L(y_i, \hat{y}^{(t-1)}) \) respectively expressed as the first and second derivatives of the loss function.

Take the partial derivative of \( j \) in (6) and set the derivative as 0. Substitute the value into (6), and the optimal objective function can be obtained as shown in equation (7).

\[
Obj^{(t)} = \frac{1}{2} \sum_{j=1}^{T} \frac{g_j^2}{H_j + \lambda} + \gamma T
\]

(4) Add the best new tree generated in the previous step to the current model.

\[
\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)
\]

(5) After all the weak regression trees are added, the iteration ends. By calling the current model and inputting single variable data for prediction, the interval prediction value can be obtained, which can be used as the reliability function value of the current variable.
2.4. Information Fusion

Information fusion of evidence theory is realized through fusion rules. Before the fusion, the reliability function value of each evidence variable, also known as the basic probability assignment function (BPA), needs to be calculated first, which satisfies Equation (9).

\[ m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \subseteq U} m(A) = 1 \]  

(9)

where, \( U \) is the identification frame. On the identification frame, Function \( m : 2^U \rightarrow [0,1] \) called a mass function. \( m(A) \) is the value of the \( A \) event's reliability function (BPA), indicates how much support a certain evidence has for the event \( A \), and \( \emptyset \) is empty set.

After the value of the body reliability function of each evidence variable is determined, the evidence theory fusion rules can be used to achieve information fusion. Let the existing four evidence body variables be \( m_1, m_2, m_3 \) and \( m_4 \), and the corresponding reliability function values are \( m_1(A_1), m_2(A_2), m_3(A_3) \) and \( m_4(A_4) \), the fusion result is \( m_{1234}(\beta) \). Represents the support degree of the film box office interval \( \beta \) after the fusion of the above four evidence variables. The fusion calculation is shown in Equation (10).

\[ m_{1234}(\beta) = m_1(A_1) \oplus m_2(A_2) \oplus m_3(A_3) \oplus m_4(A_4) = \sum_{k=1}^{3} \sum_{\beta \subseteq U, \beta \neq \emptyset} \frac{m_1(A_1)m_2(A_2)m_3(A_3)m_4(A_4)}{m_{1234}(\beta) (1-k)} \beta \]  

(10)

where \( k \in [0,1] \) represents the conflict between evidence variables. The calculation is shown in equation (11).

\[ k = \sum_{A_1 \cap A_2 \cap A_3 \cap A_4 = \emptyset} m_1(A_1)m_2(A_2)m_3(A_3)m_4(A_4) \]  

(11)

3. Method Testing

3.1. Data Collection

In this paper, two major domestic authoritative movie websites, including China Movie Ticket (http://www.cbooo.cn/) and Douban Movie (https://movie.douban.com), were used as data sources to collect a total of 379 domestic movies from 2017 to 2019 to verify the method in this paper. After deleting films with box office less than 1 million yuan and no Douban rating information, 244 films remained. The data was divided by the automatic partition function, and the partition ratio was 8:2. Among them, 195 are training films and 49 are test films.

Based on the data of the training set and the box office classification of the authoritative Chinese film rating website Douban Movie, and combined with the box office distribution of domestic films from 2017 to 2019, the film box office classification is divided into the following five categories. The cumulative box office of less than 20 million is the category of I; the cumulative box office of 20 million to 80 million is the category of II; the cumulative box office of 80 million to 200 million is the category of III; the cumulative box office of 200 million to 500 million is the category of IV; the cumulative box office of more than 500 million is the category of V.

3.2. Variables Quantization

(1) Influence of director. Considering the forgetting effect and recency effect of consumers, this paper quantified the influence of directors based on the box office of the three director-led films and their word-of-mouth, in which word-of-mouth is represented by the score of Douban. The quantization is shown in equation (12).

\[ \text{Dir} = \sum_{j=1}^{N} B_j R_j / N , \quad N = \min(n_i, 3) \]  

(12)

where, \( \text{Dir} \) is the director’s influence, \( B_j \) represents the cumulative box office of the \( j \)th film directed by the director, \( R_j \) represents the Douban score of the \( j \)th film directed by the director, and \( n_i \) is the total number of films directed by the director.

(2) Influence of screenwriter. Similar to the quantification of director’s influence, the quantification
of scriptwriter’s influence is shown in equation (13).

\[ Scr = \sum_{j=1}^{N} B_j R_j / N, \quad N = \min(n_i, 3) \]  

(13)

where, \( Scr \) is the influence of the screenwriter, \( B_j \) represents the cumulative box office of the \( j \)th film written by the screenwriter, \( R_j \) represents the Douban score of the \( j \)th film written by the screenwriter, and \( n_i \) is the total number of films written by the screenwriter.

(3) Influence of actors. Similar to the quantification of the director’s influence, the quantification of the actor’s influence is shown in equation (14).

\[ Act_m = \sum_{j=1}^{N} B_j R_j / N, \quad N = \min(n_i, 3) \]  

(14)

where, \( Act_m (m = 1, 2) \) are respectively the influence of the first leading actor and the second leading actor; \( B_j \) represents the cumulative box office of the \( j \)th film in which the actor participated; \( R_j \) represents the Douban score of the \( j \)th film in which the actor participated; and \( n_i \) is the total number of films in which the actor participated.

(4) Genre influence. According to the sample data, this paper divides the film types into 16 types, such as drama, comedy, science fiction and love. The quantification formula is shown in equation (15).

\[ Typ = \sum_{i=1}^{t} \left( \sum_{j=1}^{N} B_j / N \right) \]  

(15)

where, \( Typ \) is the influence of film type, \( j \) represents the \( j \)th film of film type \( i \), \( N \) represents the total number of films of the type \( i \) in the test set, \( B_j \) represents the cumulative box office of the \( j \)th film of the type \( i \), and \( \sum_{j=1}^{N} B_j / N \) represents the influence of the type \( i \). Considering that a film has more than one film type, the film type influence of the film is expressed as the sum of all the influences of the film type. But at the same time, the type of a movie should not be too many, Take \( t \) as the maximum of 3.

3.3. Forecasting Results

The evidences of the preparatory period were quantified and input into the trained XGBoost model, and the reliability function values of all the evidence variables can be obtained through the output of the model. According to the data performance of the training set, the model parameters are set as follows. The initial value of the learning rate is 0.1, n_estimators is 1000, max_depth is 3, Gamma is 0, Subsample is 0.8, colsample_btree is 0.8, scale_pos_weight is 1, random_state is 30.

In the results of the testing set, a film is randomly selected from five box office intervals for result display. The results are shown in table 1.

In table 2, lateral view, the method in the five testing set film performance is good, except for “Hello, Zhihua”, a reliability results outside the small difference in the second and third interval (the II forecast for reliability is 0.4, the box office forecasting for the III class reliability is 0.5, the box office is accurate prediction interval, reliability value difference is small, but not easy to estimate earnings, easy to make investors hesitate in the investment), the other four films forecast accurate results are strong convincing. Longitudinal, the influence of the first leading actor and the influence of the director has better prediction accuracy compared with other variables, and have a greater influence on the results. This also reflects the positive contribution and influence of the first leading actor and director to the box office from the side, which is in line with the objective fact. However, the influence of movie genre is the worst in the model. Since the movie genre only provides a rough summary of the movie content and does not bring more practical information about the movie quality to the audience, it performs poorly in the prediction.
Table 1. Forecast results.

| Films                  | Variables                        | Predicted results | Predicted interval | Actual interval |
|------------------------|----------------------------------|-------------------|--------------------|-----------------|
| Windfall/Value 4099    | Influence of first leading        | 1389              | I                  | I               |
|                        | first influence                  | 57327             |                    |                 |
|                        | Influence of second leading       | 1544              |                    |                 |
|                        | second influence                 | 1544              |                    |                 |
|                        | Genre influence                  |                   |                    |                 |
|                        | director influence               |                   |                    |                 |
|                        | screenwriter influence           |                   |                    |                 |
| Bureau in the Middle   | BPA (1.0,0.0,0.0,0.0,(0.4,0.4,0.1,0.1,(1.0,0.0,0.0,0.0,(0.9,0.1,0.0,0.0,(0.1,0.0,0.0,0.0,0.0),0.0)) | I | I |
| The Secret of China    | BPA (0.1,0.9,0.0,0.0,(0.2,0.8,0.0,0.0,(0.5,0.4,0.1,0.0,(0.2,0.8,0.0,0.0,(0.3,0.7,0.0,0.0,(0.1,0.1,0.0,0.0,0.0),0.0)) | II | II |
| Hello, Zhihua          | BPA (0.2,0.2,0.5,0.1,(0.1,0.3,0.2,0.3,(0.4,0.3,0.1,0.2,(0.2,0.3,0.5,0.0,(0.3,0.3,0.4,0.0,(0.1,0.4,0.5,0.0,0.0)) | III | III |
| My Dear Liar           | BPA (0.0,0.0,0.3,0.7,(0.0,0.0,0.5,0.5,(0.1,0.1,0.5,0.3,(0.0,0.0,0.3,0.7,(0.0,0.2,0.4,0.4,(0.0,0.0,0.0,2,0.8,0.0,0.0,0.0,0.0,0.0)) | IV | IV |
| The White Storm 2:     | BPA (0.0,0.0,0.0,0.1,(0.0,0.0,0.0,0.0,3,(0.0,0.0,1.0,0.0,0.7,(0.0,0.0,0.0,0.0,2,(0.0,0.0,0.0,0.0,0.0,2,(0.0,0.0,0.0,0.0,0.0,0.0)) | V | V |
| Drug Lords             | BPA (0.9,0.7,0.2,0.8,0.8,1.0))   |                   |                    |                 |

3.4. Results Comparisons

3.4.1. Forecasting Accuracy Comparisons. To illustrate the effectiveness of the method, we first performed K-fold cross validation on all samples. According to the number of film samples, K=5, and the verification results are shown in table 2.

Table 2. K-fold cross validation results.

|               | One-fold | Two-fold | Three-fold | Four-fold | Five-fold | Mean accuracy |
|---------------|----------|----------|------------|-----------|-----------|---------------|
| D-S           | 71.42%   | 75.51%   | 73.47%     | 79.59%    | 77.55%    | 75.51%        |

As shown in table 2, the accuracy of the method proposed (D-S) in this paper was over 70% in each test, the highest accuracy was 79.59%, the lowest was 71.42%, and the average accuracy was 75.51%, indicating that the method was effective to some extent.

In order to evaluate the accuracy of the proposed method, two classical classification algorithms, Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT), were selected as comparison methods to compare the prediction accuracy with the proposed method. Accuracy, MacroPrecision, MacroRecall and MacroF1 score were selected as evaluation indexes in the evaluation method of confusion matrix. The comparison results are shown in table 3.

Table 3. Forecasting accuracy comparisons.

|       | Accuracy (%) | MacroP (%) | MacroR (%) | MacroF1 (%) |
|-------|--------------|------------|------------|-------------|
| SVM   | 61.22        | 41.88      | 46.67      | 43.69       |
| GBDT  | 71.43        | 64.11      | 67.50      | 63.33       |
| D-S   | 77.55        | 74.91      | 70.00      | 66.77       |

As shown in table 3, Accuracy, MacroP, MacroR and MacroF1 of the proposed method (D-S) are
77.55%, 74.91%, 70.00% and 66.77% respectively. All the evaluation indexes are superior to the two classical classification algorithms.

3.4.2. Generalisation Ability Comparisons. In order to verify the generalization ability of the proposed method, two classical classification algorithms, SVM and GBDT, were selected as comparison methods, and receiver operation characteristic (ROC) curve was selected for evaluation.

According to the confusion matrix, the false positive rate (FPR, represents the 1-specificity, the larger the FPR, the more actual negative categories in the predicted positive categories) and the true positive rate (TPR, represents the sensitivity, the larger the TPR, the more actual positive categories in the predicted positive categories) of the sample data can be calculated. ROC curve can be constructed by taking FPR and TPR as horizontal and vertical coordinates. The closer the curve is to the upper left corner, the better the classification effect and the stronger the generalization ability of the model.

![SVM-ROC curve](image1.png)

Figure 2. SVM-ROC curve.

![GBDT-ROC curve](image2.png)

Figure 3. GBDT-ROC curve.

![D-S-ROC curve](image3.png)

Figure 4. D-S-ROC curve.

It can be seen from figures 2-4. The proposed method is superior to the other two comparison methods. The micro average is 0.95, and the macro average is 0.93, indicating that the method has good generalization ability. In addition, as we can see from figure 4, the ROC curve value of the method in the interval I, II, IV and V are all more than 0.9, and the box office interval III is 0.85, the whole curve is close to the upper left corner, indicating that the method in this paper has a good prediction effect in each box office interval.

4. Conclusion

Aiming at the difficulty in predicting the interval demand in the film preparation period caused by the few available influencing variables and the uncertainty in the prediction process, this paper proposed and verified an interval reliability classification prediction method combined XGBoost and D-S evidence theory from the perspective of data-driven and information fusion. In view of the characteristics of less variable data, XGBoost is used to complete the calculation of the value of the reliability function of the evidence variables, so that the reliability allocation is more consistent with the actual scene. At the same time, D-S evidence theory is introduced to solve the problem of uncertainty caused by single variable prediction. The case study showed that the proposed method has good credibility. Compared with the other two classical machine learning classification models, the proposed
method is superior in prediction accuracy and generalization ability.

The results of this study have the following implications. First of all, this study established the interval reliability demand prediction model in the film preparation period, which can accurately predict the box office category of films. Secondly, the importance of variables to the prediction results can be reflected in the process of model prediction, which can provide support for decision makers to select important movie features. Then, this study can effectively avoid the decision error caused by a single decision, and help the decision maker to consider comprehensively from multiple perspectives. Finally, since film is a product with short life cycle, it can be considered to expand the application scope of the method in this paper to realize the demand prediction of other products with short life cycle (such as games, etc.).

Acknowledgments
This study is supported by the National Natural Science Foundation of China (Grant No. 71672004).

References
[1] Tang Z J, Wang M Y, Yu H B, et al. 2018 Forecasting model for daily box office with improved bass model and parameter regression model Industrial Engineering and Management 23 (04) 16-22+29.
[2] Jose B M, Rodriguez J A and Dario B H 2016 Analogue-based demand forecasting of short life-cycle products: a regression approach and a comprehensive assessment International Journal of Production Research 55 (7-8) 1-15.
[3] Van B M, De R K and Wetzels M 2001 What factors determine use of quality-related marketing research information? An empirical investigation Total Quality Management 12 (4) 521-534.
[4] Ghiassi M, Lio D and Moon B 2015 Pre-production forecasting of movie revenues with a dynamic artificial neural network Expert Systems with Applications 42 (6) 3176-3193.
[5] Zhang J and Zhou S B 2014 Modeling on box-office revenue prediction of movie based on neural network Computer Application 34 (3) 742-748.
[6] Ahmed U, Waqas H and Afzal M T 2020 Pre-production box-office success quotient forecasting Soft Computing 8 (24) 6635-6653.
[7] Parimi R and Caragea D 2013 Pre-release box-office success prediction for motion pictures International Conference on Machine Learning and Data Mining in Pattern Recognition (Springer-Verlag) pp 571-585.
[8] Tang Z J and Dong S P 2020 Research on a method for forecasting the short life cycle experiential product during preparation period Control Engineering 27 (02) 226-233.
[9] Yang G J, Xu X and Zhao F P 2019 Predicting user ratings with XGBoost algorithm Data Analysis and Knowledge Discovery 3 (01) 118-126.