Prediction of fishing vessel operation mode based on Stacking model fusion

Huaichun Fu¹, Shouwei Gao²*, Yan Peng³ and Nan Zhao⁴

¹Research Institute of USV Engineering, Shanghai University, Shanghai, Shanghai, 200000, China
²Corresponding author’s e-mail: swgao@shu.edu.cn

Abstract. Due to the continuous upgrading and optimization of fishing technology and tools, and the diversification of fishing vessel operation methods, marine fishery resources are continuously depleted. Precise prediction of the operation methods of marine fishing vessels is helpful to realize effective supervision of fishing behavior of fishing vessels. In order to improve the prediction accuracy, when doing feature engineering, this paper uses a vector encoding scheme based on trajectory sequence, and uses text vectors to train the word2vec model to calculate the embedding features of each position. At present, the single method needs to be improved in terms of forecasting accuracy. This paper proposes a forecasting method based on Stacking model fusion in order to further improve the forecasting accuracy of marine fishing vessel operations. The experimental results show that the Stacking fusion model using the vector coding scheme based on the trajectory sequence has a greater improvement in prediction accuracy than a single model.

1. Introduction

The marine economy has always been an important pillar of China's economic development[1]. Marine fishery production is a core part of the marine economy. It not only enriches the food resources available to mankind, increases the income of marine economic practitioners, but also further eases the economic pressure brought about by relying on land production. Marine fishing vessels use a variety of operations to catch fish, and for different marine products, fishing vessels use different operations. According to the "Regulations on the Administration of Fishery Fishing Licenses", there are 12 types of operations approved by the fishing license, including trawl, purse seine, gill net, fishing gear, harrow thorn, trap, cage pot, ground pull net, lay net, copy net, mask. The main operation methods studied in this article are trawling, purse seine and gillnet operations. These are the three most common operations for marine fishing vessels during fishing operations, and include fishery products on all layers of the ocean. However, with the rapid growth of the number of marine fishing vessels and the rapid development of science and technology, my country's marine fishery development is also facing a series of problems[2], such as: (1) The high fishing intensity has led to the rapid depletion of marine fishery resources. (2) The state’s implementation of "closed fishing periods" and "closed fishing periods" has also led to a reduction in fishing time and scope, resulting in a reduction in fishermen’s income. (3) Trawling operation has caused great damage to the seabed ecosystem. This is because deep sea trawl operations, when dragging the net from the bottom of the sea, will also drag the large stones along the seabed along the way, causing the coral reefs to be crushed. The fish and animal that were not intended to catch will be net live, so the impact on the marine ecosystem can be devastating.
Many problems in the development of marine fisheries are related to fishing vessels. Therefore, to solve these problems, we should start with fishing vessels. In 1988, in order to monitor fishery activities, Portugal invented the vessel monitoring system (VMS), which was continuously optimized and improved. China's research and development of VMS is relatively late, but it has developed rapidly. In particular, Beidou satellite navigation and positioning system [3], which is independently developed by China, is a global positioning system with independent intellectual property rights. It's function is completely consistent with the GPS of the United States. It can provide real-time information such as ID, latitude and longitude coordinates, speed, heading, reporting time and other information of fishing vessels. Although Beidou satellite can provide the trajectories of fishing vessels, it does not include the status data of fishing vessels' operation mode, and these data can only be extracted from the log of fishing vessels [4]. There are two disadvantages of the logbook: on the one hand, there may be errors in the logbook, because it is manually input in the fishing process; on the other hand, there is a large time interval between the acquisition of the log data [5]. Literature [6] only uses speed and heading changes as the classification basis for the prediction of fishing boat operation mode, with few selections of variable factors, and no depth feature information is mined. At the same time, due to the small amount of test sample data, the experimental results may have certain errors. In addition, only using the BP neural network to process the VMS data, it is impossible to determine that this method is the best method to predict the production mode of fishing boats, because the BP neural network also has problems such as slow learning speed, easy to fall into local minimums, and overfitting.

In order to solve the above problems, this paper uses the method of machine learning to mine and analyze the location data of the Beidou equipment of the fishing vessel, predicts the production mode of the fishing vessel, and specifically determines whether the fishing vessel is engaged in trawling, purse seine or gill net operation. In order to improve the prediction accuracy, this paper uses a vector coding scheme based on trajectory sequence in feature engineering, and uses text vector to train word2vec model to calculate embedding features of each position. In addition, due to the limited prediction accuracy of single learning method, this paper proposes a prediction method based on Stacking model fusion, in order to further improve the accuracy of predicting fishing vessel operation mode. Through this data analysis and prediction method, we can effectively monitor the production and operation mode of fishing vessels, prevent excessive trawling behavior that damages the bottom marine ecosystem, and ensure the rational development and utilization of fishery resources.

2. Research contents and methods

2.1. Overall research framework

Figure 1 is the overall framework for the prediction of fishing vessel operation mode. Firstly, clean the data of the fishing vessel Beidou equipment. For example, perform missing value checking, outlier detection and elimination to meet the requirements of the learner for data; secondly, since the data set contains five original fields, such as longitude and latitude, speed, heading and reporting time, feature engineering is needed to derive the basic features; Then, the constructed features are input into Random Forest, XGBoost, LightGBM individual models and Stacking fusion model to train fishing vessel operation mode prediction; finally, the prediction results of each model are classified and output.
2.2. Exploration of fishing vessel data
The operation modes of marine fishing vessels are mainly divided into trawl operation, purse seine operation and gill net operation. The trawling operation mode is that the fishing vessel drags the fishing net forward. In the process of dragging the fishing net, the fish in the sea are "caught" and finally the net can be pulled in. The main fishing objects are shrimp, crabs and fish in bottom and middle water. Seine operation mode is that the fishing vessel will a large net to surround the fish, through a cable line to close the bottom of the net, and finally close the net. The fishing objects are mainly fish in the top water area. The gill net operation mode is to lay the long strip-shaped fishing net in the fixed water area, so that the fish can be stabbed into the mesh or entangled by the net clothes. Its characteristic is that it has a strong selectivity to the individual size of the fish, which is conducive to the protection of fish resources. The fishing objects are mainly fish scattered or concentrated in various water layers. The data set of Beidou equipment of fishing vessel used in this study is provided by Ali Yun Tianchi competition. The field name of the data set contains information such as fishing vessel ID, longitude and latitude coordinates, speed, course, reporting time, operation mode, etc. In order to observe the trajectory characteristics of fishing vessels with different operation modes, a trawler with ID 20153, a purse-seine fishing vessel with ID 2001 and a barbed-net fishing vessel with ID 20338 were randomly selected from the data set for visualization in a two-dimensional plane, as shown in Fig. 2A, Fig. 2B and Fig. 2C respectively:
According to the distribution map of trawl trajectory characteristics, trawl fishing vessels need to drag fishing gear for fishing, so they are in a straight state most of the time, with fewer turns, and their trajectories generally do not form a closed state; According to the distribution map of the characteristics of the purse seine trajectory, since the purse seine fishing vessel usually needs to circle during the fishing process, the turning range is large, and its trajectory will also form a closed state. According to the distribution map of gill-net trajectory characteristics, gill-net fishing usually only needs to place the fishing net somewhere and wait, so when gill-net fishing, the fishing vessel tends to be stationary and sometimes may move slowly with the sea wind or current, with a smaller moving range.

2.3. Preprocessing

2.3.1. Missing Value Check. This paper uses Python language for data preprocessing and Pandas database for missing values checking. Processing steps are: (1) Importing the required tools; (2) Code reading the overall situation of the data of the fishing vessel Beidou equipment; (3) Viewing the missing attributes of the dataset. The final operation results are shown in Table 1. The operation results show that the data of Beidou equipment is complete and there is no missing data.

| id | lat | lon | sd | fx | time | label |
|----|-----|-----|----|----|------|-------|
| 0  | 0   | 0   | 0  | 0  | 0    | 0     |

2.3.2. Outlier handling. A sample of a fishing vessel with ID 25536 is selected from the dataset and its trajectory is displayed on a two-dimensional plane, as shown in Figure 3A:
Figure 3. Distribution of fishing vessel track points

It can be seen from the figure that there are obvious abnormal outlier trajectory points in the trajectory of the fishing vessel. DBSCAN (Density-Based Spatial Clustering of Applications with Noise, density-based clustering method with noise) is a density-based spatial clustering algorithm [7]. The algorithm is a density-based clustering algorithm, which requires that the number of objects (points or other spatial objects) contained in a certain area in the space is not less than a specified threshold. The advantage of the DBSCAN algorithm is that the clustering speed is very fast, and noise points can be effectively removed.

The basic principle of this algorithm is to scan the entire dataset, check the $Eps$ neighborhood of each point to search for clusters, and create a cluster with $p$ as the core object if the $Eps$ neighborhood of point $p$ contains more points than the threshold $Minpts$; then, DBSCAN iteratively gathers the objects that are directly density reachable from these core objects. This process may involve the merging of some density reachable clusters; finally, when no new points are added to any clusters, the process ends. Data points that are not included in any cluster constitute abnormal points.

The meaning of the parameters in the algorithm is as follows:

1. **Eps neighborhood**

   In data set $D$, for any data point $P_i \in D$, the set of all points whose distance is less than or equal to $Eps$ with it as the center.

   $$Eps(p) = \{ q \in D \mid dist(p, q) \leq eps \}$$

2. **Threshold $Minpts$**

   In data set $D$, for any data point $P_i \in D$, the number of data points in the neighborhood of $Eps$ is set to $k$. When $k \geq Minpts$, the neighborhood of $Eps$ is classified into a class, where $Minpts$ is the threshold of the neighborhood.

   DBSCAN algorithm is used to detect and remove outliers from trajectories with large active range. The final fishing vessel sample trajectories are shown in Figure 3B. It is clear that the anomalous track point on the left side of the graph has been detected and deleted. Finally, the whole dataset can be optimized by using this algorithm to detect and delete abnormal outlier trajectory points for each fishing vessel sample in the dataset.
2.4. Feature engineering

2.4.1. General feature statistics. The original Beidou satellite equipment dataset only has five original fields, such as longitude and latitude, speed, heading, reporting time, and so on. The features that can be used directly are few, and the effect of predicting fishing vessel operation mode is not ideal. For this reason, on the basis of the original features, the original features are derived by using the statistical analysis method. For example, location-based feature derivation, statistical longitude and latitude mean, variance, median, quantile, skewness, kurtosis, and so on; speed-based feature derivation, statistical speed mean, variance, skewness, kurtosis, and so on; direction-based feature derivation, first-order difference, variance of statistical direction, and so on.

2.4.2. Use word2vec to calculate embedding features. Since the trajectory characteristics of the three fishing vessel operation methods are different, in addition to the conventional statistical characteristics, this article also uses a vector coding scheme based on the trajectory sequence for the trajectory characteristics of the fishing vessel operation methods. Use the text vector training word2vec model to calculate the embedding features of each position [8-9], so as to train the model and improve the prediction accuracy of the model.

In the field of natural language processing, the word2vec model is the preferred model for generating word vectors. Word2vec model was proposed by Mikolov [10] et al. after optimizing the structure of Neural Network Language Model (NNLM). Simply put, this model maps words or phrases in the vocabulary to a real vector space so that the similarity between words can be better measured by the word vector. The word vector generated by word2vec model contains two kinds of information: one is that it contains the information expressed by the word itself; the other is that the information of the semantic relationship between words can be determined by the position of the word vector in the vector space. In addition, the addition and subtraction of word vectors can add or subtract their own semantics, which is word offset technology. Finally, the distance between two word vectors generated by the word2vec model in vector space can also measure the similarity between two words. The closer the words are, the more semantically similar they are. Therefore, we can calculate the degree of similarity between two words by calculating their cosine similarity or Euclidean distance.

The specific process of using the word2vec model to calculate the vector encoding scheme of word vectors is shown in Figure 4:

![Figure 4. Vector coding scheme based on trajectory sequence](image)

In the input layer, the longitude and latitude sequences of fishing vessels are input; in the data processing layer, Geohash code [11] is used, where latitude and longitude are not used directly to enhance the generalization ability of features; in the feature coding layer, a special feature coding method is used, that is, DNN code based on word2vec (DNN is a deep neural network); in the output layer, the feature vectors of fishing vessel tracks are output. Specifically, the entire word vector feature generation process is as follows: first use the Geohash algorithm to encode the latitude and longitude, then convert the position of each fishing vessel at each time into a text vector, and then use the text vector to train the word2vec model, then use the word2vec model obtained to calculate the
embedding feature of each position, and finally take the average value of the embedding feature of each fishing vessel at all times.

2.4.3. Feature selection. The features finally constructed above may be highly correlated, which is not conducive to model training. Therefore, it is necessary to perform correlation analysis on all these constructed features, and eliminate highly correlated variables to improve the speed and accuracy of model training. When making feature selection, there are many methods that can be used, the common ones are Pearson correlation coefficient, distance correlation coefficient, mutual information and maximum information coefficient, Pearson chi-square test, etc.

The feature selection method selected in this paper is Pearson correlation coefficient [12], which is a simple and useful feature selection method, and the effect of feature selection is also good. Pearson correlation coefficient values are between -1 and 1. It reflects the linear correlation between the feature and the predicted value, and describes the trend of the two sets of linear data changing simultaneously. Among them, if the feature and the predicted value are absolutely negatively correlated, the Pearson value is -1; if there is no linear correlation between the feature and the predicted value, the Pearson value is 0; if the feature and the predicted value are absolutely positively correlated, the Pearson value is 1. And the formula is as follows:

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$  \hspace{1cm} (2)

In formula (2), \(\text{cov}(X,Y)\) represents the covariance of two variables, and its calculation formula is as follows:

$$\text{cov}(X,Y) = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$  \hspace{1cm} (3)

In formula (2), \(\sigma_X\) represents the standard deviation of variable \(X\), and the calculation formula is as follows:

$$\sigma_X = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$  \hspace{1cm} (4)

In formula (2), \(\sigma_Y\) represents the standard deviation of variable \(Y\), and the calculation formula is as follows:

$$\sigma_Y = \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$  \hspace{1cm} (5)

After feature selection, combined with the LightGBM algorithm, the top 20 features ranked by feature importance are shown in Figure 5:
Figure 5. Feature importance ranking

It can be seen from Figure 5 that the feature vector calculated by word2vec has a great contribution to the prediction accuracy of the model. Therefore, the vector coding scheme based on trajectory sequence proposed in this paper is very helpful to the improvement of model prediction accuracy.

2.5. Stacking model construction

2.5.1. Stacking algorithm principle. Ensemble learning is not a single learning method, but a combination of multiple machine learning techniques to construct a new predictive model to complete the learning goal. That is, through integrated learning, the effect of reducing variance, bias or improving prediction accuracy can be achieved. Generally speaking, ensemble learning can be divided into three categories: Bagging to reduce variance, Boosting to reduce bias, and Stacking to improve prediction results. They use parallel, serial, and tree row calculation methods, respectively. The Stacking fusion model used in this article is an ensemble learning method using tree-line calculation method. The Stacking method can integrate the advantages of different learning models, that is, the advantages of a hundred families. Regardless of the improvement of model prediction accuracy, generalization ability or robustness, the Stacking fusion model is more effective than a single model.

The structure of the Stacking model is generally two-layered. In Zhou Zhihua [13]'s "Machine Learning" book, the first-layer learner is generally called the primary learner, and the second-layer learner is generally called the secondary learner. First, use the original data set to train multiple different primary learners in the first layer, then construct the prediction results of the primary learners as new input features, and then train the secondary learners to get a final prediction result. However, in this process, over-fitting may occur. In order to reduce the risk of over-fitting of the model, K-fold cross-validation [14-15] is usually used to train the primary learner. Here is a 5-fold cross-validation Stacking model as an example, the algorithm principle is shown in Figure 6:

![Image of the Stacking algorithm principle](image-url)
First, the original data set is divided into the original training set (Training Data) and the original test set (Test Data). Then, perform 5-fold cross-validation on the primary learner: divide the training set into 5 equal parts randomly, take out one of them as test data, and the other 4 as training data. Combine each cross-validated prediction result as the input feature of the secondary learner. In addition, in the 5-fold cross-validation project, it is also necessary to predict the original test set Test Data. This process will generate 5 prediction result data sets. For this part of the data, the average is taken as the test set data of the secondary learner. Then, the secondary learner model is trained by using the output result of the primary learner as training data. Finally, the test data of the secondary learner will be predicted to obtain the final prediction result.

2.5.2. Stacking model construction. The model in this article uses Stacking to establish a two-layer architecture: the first layer combines different primary learners, including Random Forest, XGBoost, LightGBM and GBDT; the second layer uses SVM as secondary learner. The second-layer learner will use the predicted results of the first-layer as features and predict the final result. In the model building process, in order to reduce over-fitting, five-fold cross-validation was used. The overall architecture of the Stacking model is shown in Figure 7:

![Stacking model overall architecture](image)

Figure 7. Stacking model overall architecture

During the model building process, 8166 pieces of sample data were randomly divided into 70%, that is, 5716 pieces of data were used as the training set for model construction, and the remaining 30% of 2450 pieces of data were used as the test set to test the model.

3. Experimental results and analysis
The data set used in this article comes from the Aliyun Tianchi competition. It contains a total of 8166 sample data, including 5 related parameters such as latitude and longitude, speed, heading, and reporting time. The predicted target operation methods of fishing vessels are trawling, purse seine and gillnet operations, which are typical three-category problems. All these features after the above feature selection are input into GBDT, Random Forest, XGBoost, LightGBM individual models and Stacking
model are trained, and finally predictive analysis is performed. This article uses Python language for data processing, model building and model evaluation.

3.1. Model evaluation index

This article studies a typical three-classification problem. The following explains it through the confusion matrix shown in Figure 8:

![Confusion matrix for three classification problems](image)

The horizontal axis represents actual negative class, actual neutral class, and actual positive class (real result), and the vertical axis represents predicted negative class, predicted neutral class, and predicted positive class (predicted result). The first number of T00 represents the true result, that is, the label value, and the second number represents the predicted category, that is, the predicted value. T00 is the correct number of trawl net samples, T11 is the correct number of purse seine samples, and T22 is the correct number of gillnet samples. F01 is the number of trawl samples that were incorrectly classified as purse seine samples, and F02 is the number of trawl samples that were incorrectly classified as gillnet samples. F10 is the number of seine samples that were incorrectly classified as trawl samples, and F12 is the number of seine samples that were incorrectly classified as gillnet samples. F20 is the number of gillnet samples incorrectly classified as trawl samples, and F21 is the number of gillnet samples incorrectly classified as purse seine samples.

This article uses Accuracy, Precision, Recall, and the average of the F1 values of the three categories (Score) to measure the performance of the model.

The Accuracy formula is described as:

\[
Accuracy = \frac{T00 + T11 + T22}{N}
\]  

(6)

Note: N represents the total number of samples.

The Precision formula is described as:

\[
Precision = \frac{P0 + P1 + P2}{3}
\]  

(7)

Note: \(P0 = \frac{T00}{T00 + F10 + F20}\), \(P1 = \frac{T11}{F01 + T11 + F21}\), \(P2 = \frac{T22}{F02 + F12 + T22}\)

The Recall formula is described as:

\[
Recall = \frac{R0 + R1 + R2}{3}
\]  

(8)

Note: \(R0 = \frac{T00}{T00 + F01 + F02}\), \(R1 = \frac{T11}{F10 + T11 + F12}\), \(R2 = \frac{T22}{T20 + F21 + T22}\)

The Score value formula is described as:

\[
Score = \frac{F1_{Trawl} + F1_{Seine} + F1_{Gill Net}}{3}
\]  

(9)

Note: \(F1_{Trawl} = \frac{2 \times P0 \times R0}{P0 + R0}\), \(F1_{Seine} = \frac{2 \times P1 \times R1}{P1 + R1}\), \(F1_{Gill Net} = \frac{2 \times P2 \times R2}{P2 + R2}\)
3.2. Model performance comparison

The single model in this article uses GBDT, Random Forest, XGBoost and LightGBM. The Stacking model uses a two-layer architecture. The primary learner uses Random Forest, XGBoost and LightGBM, and GBDT; the secondary learner uses SVM. The following compares the accuracy rate, precision rate, recall rate, and Score value between the single model and the Stacking model. The results are shown in Table 2 and Figure 9.

| Model name    | Accuracy | Precision | Recall | Score  |
|---------------|----------|-----------|--------|--------|
| GBDT          | 90.16%   | 90.16%    | 85.59% | 87.34% |
| Random Forest | 91.88%   | 92.16%    | 88.22% | 89.82% |
| XGBoost       | 92.73%   | 92.36%    | 89.99% | 91.03% |
| LightGBM      | 93.06%   | 92.53%    | 90.26% | 91.26% |
| Stacking      | 94.08%   | 93.7%     | 91.66% | 92.58% |

It can be seen from Table 2 and Figure 9 that the classification prediction accuracy of a single model is relatively high, but after using the Stacking model, the classification prediction accuracy can be further improved, and other evaluation indicators have also been improved to a certain extent. The standard accuracy rate and Score value of the main evaluation index increased to 94.08% and 92.58% respectively. This shows that the stacking-based fusion model achieves a better prediction effect for the prediction method of fishing boat operation mode, and the generalization of the model is higher, which has laid a solid foundation for the practical application of the model.

4. Conclusions

This paper studies the problem of forecasting the operation mode of marine fishing vessels. Because the trajectory characteristics of the three fishing vessel operation modes are different, in addition to the conventional statistical characteristics, this article also uses a vector coding scheme based on the trajectory sequence for the trajectory characteristics of the fishing vessel operation mode. In this scheme, the text vector is used to train the word2vec model to calculate the embedding features of each position for the algorithm model training. Aiming at the low prediction accuracy of a single model, a fishing vessel operation mode prediction algorithm based on Stacking model fusion is proposed. Combining different classification models through a two-layer stacking method and using five-fold cross-validation to prevent over-fitting. This fusion algorithm that combines the advantages of multiple models further improves the prediction accuracy. It can be concluded that the fishing vessel operation mode prediction scheme based on the integration of the Stacking model can
effectively identify the fishing vessel operation mode, ensure the rational development and utilization of fishery resources, and promote the healthy and sustainable development of the marine economy.

Acknowledgments
Thanks to Research Institute of USV Engineering of Shanghai University for its support. This work was supported by the National Natural Science Foundation of China under Grants of U1813217.

References
[1] Song X. Analysis on Output Efficiency of Marine Fishery in China's Coastal Provinces and Cities[J]. Journal of Coastal Research, 2020, 115(sp1):81.
[2] Su S, Tang Y, Chang B, et al. Evolution of marine fisheries management in China from 1949 to 2019: How did China get here and where does China go next?[J]. Fish and Fisheries, 2020, 21(2).
[3] Zhou W, Hou J, Fan W, et al. A location-based marine fishery information service based on BeiDou navigation satellite system[J]. Journal of Intelligent and Fuzzy Systems, 2019, 37(4):1-8.
[4] Yang Wang, Yingbin Wang, Ji Zheng. Analyses of trawling track and fishing activity based on the data of vessel monitoring system (VMS): A case study of the single otter trawl vessels in the Zhoushan fishing ground[J]. Journal of Ocean University of China, 2015, 14(1).
[5] Youen Vermard, Etienne Rivot, Stéphanie Mahévas, Paul Marchal, Didier Gascuel. Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian Hidden Markov Models[J]. Ecological Modelling, 2010, 221(15).
[6] Qiaoling Z, Wei F, Shengmao Z, et al. Identification of fishing type from VMS data based on artificial neural network[J]. South China Fisheries ence, 2016.
[7] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, 1996, pp. 226–231.
[8] Dong H, Bo X U, Kan X U, et al. Short text clustering based on word embeddings and EMD[J]. Journal of Shandong University (Natural ence), 2017.
[9] Kolesnikova O, Gelbukh A. A study of lexical function detection with word2vec and supervised machine learning[J]. Journal of Intelligent and Fuzzy Systems, 2020(2):1-8.
[10] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer ence, 2013.
[11] Xiang W. An Efficient Location Privacy Preserving Model based on Geohash[C]// 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC). 2019.
[12] Jiang-Yong Z. Study on Combination Forecast of Electricity Consumption Based on Pearson Correlation Coefficient Criterion[J]. Journal of Yangzhou Polytechnic College, 2016.
[13] Zhou, Z. (2016) Machine Learning[M]. Tsinghua University Press, Beijing.
[14] Wong T T, Yeh P Y. Reliable Accuracy Estimates from k-fold Cross Validation[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, PP(99):1-1.
[15] Jinbo H, Xitao F. Evaluating the Performance of the K-fold Cross-Validation Approach for Model Selection in Growth Mixture Modeling[J]. Structural Equation Modeling A Multidisciplinary Journal, 2018:1-14.