International Journal of Cognitive Informatics and Natural Intelligence
Volume 16 • Issue 1

Recognition of Air Passengers' Willingness to Pay for Seat Selection for Imbalanced Data Based on Improved XGBoost

Baiyu Hong, Huzhou University, China
Xiaolong Ma, Huzhou University, China*
Weining Tang, Huzhou University, China
Zhangguo Shen, Huzhou University, China

ABSTRACT

Passenger-paid seat selection is one of the important sources of ancillary revenue for airlines, and machine learning-based willingness-to-pay identification is of great practicality for airlines to accurately tap potential willing passengers. However, affected by periodic statistical errors, air passenger order data often has some problems such as high noise, high latitude, and unbalanced category. In view of this, this paper proposes a method for identifying air passengers' willingness to pay for seat selection based on improved XGBoost, which is improved and integrated from three stages: data, feature, and algorithm. The feasibility of the proposed multi-stage improved integration method is verified by real airline passenger dataset, and the experimental results show that the proposed improved method has better classification effect when compared with the classical six imbalance classification models, which provides a basis for accurate marketing of airline paid seat selection programs.

KEYWORDS

Air Passenger, Imbalanced Data, Paid Seat Selection, Whale Optimization Algorithm, XGBoost

1. INTRODUCTION

Affected by the global economic environment and market competition, airline's main ticket business revenue has gradually decreased with the continuous decline in ticket prices, and many airlines have begun to generate revenue by increasing the added value of products and developing ancillary services to ease the financial pressure on their operations. In addition, coupled with the huge impact of the novel coronavirus pneumonia on global air passenger demand, better performance of ancillary revenues can help airlines survive the epidemic crisis to some extent. The "paid seat selection service", one of the new domestic add-on services, has brought the airlines considerable profits due to its almost zero marginal expenditure. Since ancillary products and services are optional, so it is very important for airlines to understand passengers' willingness to pay and let more passengers choose this service.

DOI: 10.4018/IJCINI.312249 *Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Over the years, A great deal of exploration has been done in the identification of customers' willingness to pay for various ancillary services, and machine learning-based willingness identification methods have been found to be more advantageous than traditional statistical methods (Jing et al., 2021; Maliah & Shani, 2021). By studying and analyzing the travel purpose of the ticket-purchased passengers, mining the behavioral characteristics of the known paid seat-selecting passengers and constructing their behavioral models (Borisyak et al., 2020; J. Pang, Chen, Li, Xu, & Lin, 2021), it is possible to identify the passengers who may have similar willingness and ability from the total number of passengers, so as to achieve increased revenue from accurate marketing at a lower cost.

The current research on the travel behavior of civil aviation passengers is divided into two main directions: passenger behavior segmentation (Pan & Truong, 2021) and passenger value calculation (Nakahara & Yada, 2011). Most airlines subdivide passengers according to the fare of the ticket they purchased or the accumulated mileage distance. The RFM model proposed by marketing expert Bob Stone can be used to quantify customer value, many scholars (Wu et al., 2020; Wu et al., 2021; Zong & Xing, 2021) use RFM and improved k-means clustering to divide passenger groups. However, such segmentation method only discovers the value of passengers and does not point out the behavioral characteristics of passenger groups. The three attributes of traditional RFM model do not fully reflect the passenger behavior preferences, so on this basis, the LRFMC model and the TCSDG model have been successively evolved. Combining the improved model with classical machine learning classification algorithms such as SVM, KNN, GBDT and NN, etc., it is widely used in passenger behavior prediction (S. Q. Pang & Liu, 2011), personalized recommendation (Tao, 2020), flight delay prediction (Jiang, Liu, Liu, & Song, 2020) and other aviation fields in business. However, experimental studies on identifying passengers' willingness to choose a seat for a fee are still limited.

The biggest challenge in the field of air passengers' pay-for-seat willingness identification is its noisy, latitudinal, and uneven distribution due to environmental and recording influences when collecting data. Therefore, this paper proposes an improved XGBoost-based method for identifying airline passengers' willingness to take paid seats based on high-dimensional imbalance data, which conducts integrated classification prediction from three stages of data, features and algorithm.

The main contributions made in this paper are shown below:

1. **Data preprocessing:** Random undersampling of the majority class samples and generation of new samples to supplement the minority class samples using CGAN, so as to solve the serious imbalance of passenger order data.
2. **Feature selection:** Then combining chaos theory, introducing nonlinear convergence factors and adaptive weights, improving the whale optimization algorithm based on the opposing learning strategy to find the optimal feature subset and reduce the dimensionality of the traveler data set.
3. **Algorithm improvement:** A new gradient harmonizing mechanism (GHM) is introduced to improve the loss function of XGBoost, and the genetic algorithm is used to optimize the parameters of XGBoost to obtain the final willingness recognition model.

The experimental results with real airline passenger data show that the improved classification model of this paper has a better classification effect than the other six representative imbalance classification models SMOTEBoost (Chawla, Lazarevic, Hall, & Bowyer, 2003), RUSBoost (Seiffert, Khoshgoftaar, Van Hulse, & Napolitano, 2010), CUSBoost (Rayhan, Ahmed, Mahbub, Muhammod, & Farid, 2017), EasyEnsemble (X. Y. Liu, 2006), SMOTE_SVM (Nguyen, Cooper, & Kamei, 2011) and Balanced Random Forest (Chao & Breiman, 2004). This study fully proves the superiority of multi-stage improved ensemble method in the field of passengers' willingness to pay for seat selection recognition compared with a single model. The proposed improvement provides some certain reference for follow-up related research.

2. RELATED WORK

2.1 Airline Database

Paid seat selection has gained much attention in recent years as one of the new additional services. At present, all major airlines basically have their own database systems and have accumulated a large amount of passenger data. How to identify the most profitable and potentially willing passengers from the huge amount of passenger data is an urgent problem for all airlines to solve. Many scholars have also conducted relevant studies on this topic. Liou et al. (Liou & Tzeng, 2010) used the Dominance-based Rough Set Approach (DRSA) to provide a set of rules for determining customer attitudes and loyalty, which can help airlines to identify high-value customers and enhance their attractiveness to potential users. Then developing strategies to retain high-value customers and acquire new ones. Li et al. (L. Li, 2017) used the "Seat Reservation" information provided by Air China to analyze the booked passengers in six aspects and multiple dimensions to build a positioning model to help identify potential users for the "Seat Reservation" value-added service, and help airlines develop targeted strategies. Lu et al. (Lu, 2018) collected the historical operation records of frequent flyers of China Southern Airlines from 2012 to 2016, and used stacking to fuse individual models to predict lost passengers and analyze their behavioral characteristics, providing an effective method to identify passenger loyalty. Rouncivell et al (Rouncivell, Timmis, & Ison, 2018) set up a 14-question survey via an online survey platform on willingness to pay for seat selection on UK flights unrelated to the ticket, providing an evidence base for the development of revenue management and the marketing of seat selection fees as an ancillary product. Yu et al. (Yu, 2019) used an improved RFM-P model to segment airline customers, and then analyzed the factors influencing value-added service purchase decisions for the segmented customer groups to investigate which factors can facilitate airline passengers' purchase of value-added services. Zhao et al. (Zhao, 2021) used the K-means++ to cluster passengers into three groups: "low-value passengers", "medium-value passengers", and "high-value passengers", and constructed an improved XGBoost-based passenger churn prediction model with churn rules, enterprises can get the loss of passengers with different values and formulate differentiated retention strategies. While many of the above scholars have provided many experiences and results on airline value-added services, research in the area of willingness identification specifically related to paid seat selection is still very limited.

The data used in this experiment is the real booking data of an airline for three years from January 2018 to December 2020, which comes from the China Student Service Outsourcing Innovation and Entrepreneurship Competition (The official website of the competition is http://www.fwwb.org.cn/news/show/314, accessed on 22 June 2022). And the source of the data is related to the airline's precise marketing activities.

There are a total of 23433 records in the experimental data set, and each record includes 656 passenger information variables and 1 target variable. The passenger information variables can be divided into three parts: passenger basic information, Passenger information on a certain flight and passenger flight statistics during the period, including 523 numerical characteristics and 133 non-numerical characteristics. The target variables are expressed by 0 and 1, where 0 represents the unpaid seat selection sample and is the majority class, and 1 represents the paid seat selection sample and is a minority class. There are 1475 minority samples and 21957 majority samples, with an unbalanced ratio of 1:15. The specific information is shown in Table 1.

As can be seen from Table 1, passenger seat selection willingness recognition is to classify the total number of passengers by mining the characteristic information of paid passengers to achieve precision marketing, which is essentially a binary classification problem. However, in the actual sample of air passengers, the number of passengers who do not pay for seat selection is usually far more than that of passengers who have paid for seat selection. and this unevenly distributed data is also widely available in other practical applications, such as disease diagnosis (Devarriya, Gulati, Mansharamani, Sakalle, & Bhardwaj, 2020), fraud identification (Y. Liu, Yang, K., 2021), text

Table 1. Experimental data description

Data Category	Data Sources	Sample Structure	Characteristic Attribute
Air passenger paid seat selection	An airline	Minority samples: 1475	Total number of features: 657
		Majority samples: 21975	Data types: numeric and category
		unbalance degree: 1:15	Preprocessing: CGAN and IWOA

classification (Elnagar, Al-Debsi, & Einea, 2020), machine fault detection (X. Q. Li, Jiang, Liu, Zhang, & Xu, 2021), spam filtering (Dedeturk & Akay, 2020), etc. When constructing the willingness recognition model, traditional classifiers prioritize the classification accuracy of the majority class samples (unpaid seat selection) in order to ensure the overall performance of the model training, resulting in some minority class samples (paid seat selection) being classified incorrectly. And in the willingness recognition problem, the portrait information of paid passengers has a higher value, that is, misidentification of a willing passenger often causes more losses to airlines than misclassification of an unwilling passenger. Therefore, how to detect more paid minority class passengers and improve the classification effect of the model on imbalanced passenger data is the focus of research in the field of air passengers' willingness to pay for seat selection identification.

2.2 Unbalanced Data

In view of the shortcomings of the above-mentioned unbalanced data in the classification problem, many scholars have carried out in-depth research and analysis, and have put forward many solutions. There are mainly three types of solutions: data preprocessing, feature selection and algorithm improvement.

1. **Based on data preprocessing:** At the data level, data distribution is mainly adjusted by resampling or grouping the data to weaken its distribution imbalance. Among them, resampling can be divided into oversampling for the minority classes, undersampling for the majority classes, and hybrid sampling combining the two sampling methods. Oversampling is to synthesize new minority class samples by some strategies to balance the number of positive and negative samples. The classical methods include SMOTE (Sindhu & George, 2022), SMOTE-Borderline (Han, Wang, & Mao, 2005), ADASYN (He, Yang, Garcia, & Li, 2008), and GAN-based methods (Lee & Park, 2021). However, the samples generated by artificial random replication may bring noise and reduce the classification accuracy of the model. At the same time, some redundant samples are easy to cause model overfitting and increase the training time of the model. Undersampling is to screen out some representative negative samples by randomly deleting some majority class samples, so that the data set tends to be balanced. Undersampling algorithms include density-based undersampling (Cui, Cao, & Liang, 2020) and weight-based undersampling (Xiong, Wang, & Deng, 2016). However, undersampling may lose some important feature information in the process of eliminating most class samples, thus affecting the accuracy of model classification. Hybrid sampling is to rebalance the data by combining undersampling and oversampling, which can make up for the defects caused by a single method to some extent. Zhu et al. (Zhu, Yan, Zhang, & Zhang, 2020) proposed an evolutionary hybrid sampling technique that made the decision boundary between the majority class and the minority class samples is more visible, and random oversampling combined with imbalance ratio and classification performance was used for minority class samples, which reduced the risk of overfitting and enhanced the learning performance of the classifier. Gao et al. (Gao et al., 2020) divided the data space into four different regions according to the proportion of majority samples in the minority class neighborhood, different sampling methods in different regions can better balance the data distribution. Li et al. (Dongdong

et al., 2021) distinguished the importance of training samples by calculating the information entropy value in the undersampling process, and oversampling only expanded the positive data to the size of each negative sample subset, which alleviated the data overlap problem caused by artificial synthetic samples. Low et al. (Low, Cheah, & You, 2021) proposed an imbalanced class processing method combining gradient boosting algorithm and hybrid sampling. Although this method increased the possibility of model overfitting, it greatly improved the classification accuracy of minority samples, which was more advantageous than the single sampling method.

2. **Based on feature selection:** The passenger order data collected from the actual air transportation industry usually has high dimensions, and the unbalanced category of high-dimensional data set also leads to the uneven distribution of feature attributes. The redundant irrelevant features will increase the training time of the sample, resulting in overfitting of the model. Therefore, feature selection seeks the optimal feature subset which contributes most to the model and can reflect the characteristics of unbalanced data from the original high-dimensional data to reduce the complexity of the model and improve the classification performance. Common feature selection methods are mainly divided into three categories: filter, wrapper (Kohavi & John, 1997) and embedding. When facing high-dimensional imbalance problems, many scholars have also made relevant improvements to the traditional methods. Hosseini et al. (Hosseini & Moattar, 2019) combined symmetric uncertainty and binary interaction information to identify candidate features, then formed candidate feature subsets through the multivariate interaction information method and selected the best candidate feature subset, which reduced the data dimension thus improves the accuracy and f1 value. Chen et al. (H. M. Chen, Li, Fan, & Luo, 2019) proposed a feature selection method for imbalanced data based on neighborhood rough set theory, which fully considered the fuzzy distribution of class and class boundary, and verified the effectiveness of the method in binary and multi-class data. Shahee et al. (Shahee & Ananthakumar, 2020) used the effective complex distance measure to properly select the iteratively modified features, so as to obtain the final feature ranking and the unbalanced features between and within the classes are combined. Sharifai et al. (Sharifai & Zainol, 2021) combined the feature ranking of the six filters to select the feature set that exceeds the set threshold, then searched the feature space and mine high-quality features to enhance the capability to predict minority classes. Kim et al. (Kim, Kang, & Sohn, 2021) applied the ensemble learning paradigm to the feature evaluation process through the feature evaluation scheme based on filtering method, which accurately recognized the features with good robustness and greatly reduces the calculation time.
3. **Based on algorithm improvement:** Traditional classification algorithms such as decision tree, SVM, neural network is mostly trained based on the premise of data balance, which also leads to the classifier biased towards majority class samples. Therefore, the algorithm level mainly deals with unbalanced data by improving traditional classification algorithms, where cost-sensitive learning (Thai-Nghe, Gantner, Member, IEEE, & Schmidt-Thieme, 2010) and ensemble learning method (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012) have outstanding effects. Cost-sensitive learning improves the detection rate by giving higher misclassification costs to minority samples, and reduces the impact of misclassification, Ensemble learning is to combine the learned multiple sub-classifiers into a strong classifier, which is usually combined with other imbalanced data classification processing methods to comprehensively improve the classification effect. Wong et al. (Wong, Seng, & Wong, 2020) found the optimal value of cost vector by randomly undersampling in hidden layer of deep neural network and extracting feature layer by layer, and good results were achieved in the binary classification of unbalanced business field. Vong et al. (Vong & Du, 2020) proposed a new sequential ensemble learning framework to solve the multi-classification problem of highly imbalanced data. The combination method was formulated for the extreme learning machine to weaken the sensitivity to IR and improved the average classification accuracy. Alves Ribeiro et al. (Ribeiro & Reynoso-Meza, 2020) proposed an ensemble learning method based on different multi-objective optimization design to improve

the level imbalance problem. It not only improved the integration creation on fewer instances and feature training, but also improved the performance of RF and RUSBoost integration. Du et al. (Hongle, Yan, Gang, Lin, & Chen, 2021) reduced the impact of unbalanced data stream and improved the classification accuracy by dynamically calculating the cost of misclassification, sampling probability and the weight of base classifier, especially suitable for the recognition of unknown intrusion behavior in network intrusion detection. Razavi-Far et al. (Razavi-Far, Farajzadeh-Zanajni, Wang, Saif, & Chakrabarti, 2021) used missing data imputation techniques to generate new samples, it proved that the performance of the proposed oversampling combined bagging strategy based on multiple imputation was significantly better than that of the commonly used class imbalance learning method. Chen et al. (Z. Chen, Duan, Kang, & Qiu, 2021) combined ensemble learning with edge-based undersampling and diversity-enhanced oversampling to better solve the problem of poor performance caused by single data-level method, the disadvantage was that the integration scheme was greatly affected by the parameters.

As a summary, the above methods can solve the problem of imbalanced data classification to a certain extent, but there are still some limitations in the actual application of recognition of the willingness of air passengers to pay for seat selection.

3. THE PROPOSED METHOD

Aiming the problem that the traditional classifiers have a low recognition rate of paid samples on high-dimensional unbalanced passenger data, this paper will further improve XGBoost, which has excellent performance in the field of binary classification of passenger willingness to pay, in three stages: Data preprocessing, combining random under-sampling and CGAN to balance passenger samples; Feature selection, using the improved whale optimization algorithm to filter out the feature subset that contributes the most to the willingness recognition model; Algorithm improvement, a new gradient harmonizing mechanism (GHM) is introduced to improve the loss function of XGBoost, and genetic algorithm is used to optimize the parameters of XGBoost, then the final model is obtained to identify passengers' willingness to pay for seat selection. The overall framework of the algorithm is shown in Figure 1.

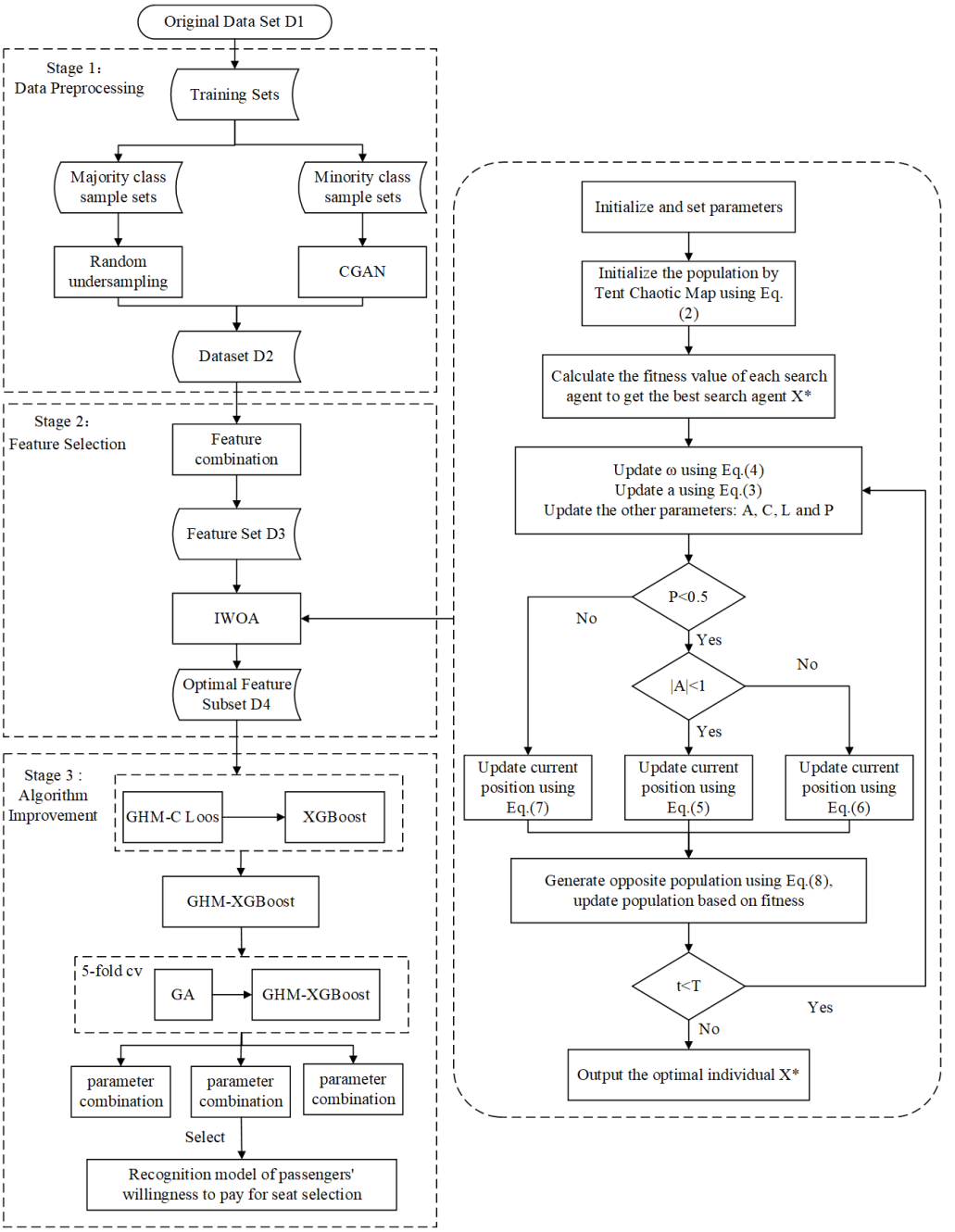
3.1 Data Preprocessing Based on Hybrid Sampling

Due to the single sampling method has its own defects, in order to improve the quality of passenger samples after resampling and to ensure that the samples are as diverse as possible (non-overlapping or less), a hybrid sampling strategy based on random undersampling and CGAN is proposed. The specific steps of the method are:

Step 1: Undersampling of negative samples. The undersampling of negative samples mainly includes two parts. Firstly, the noise samples in negative samples are removed to reduce the interference to the classification. Secondly, since the number of negative class samples is much larger than positive class samples in the overlap region, making the positive class boundary samples in the overlap region often harder to identify, the proportion of negative classes in the overlap region should be reduced. In this paper, the majority and minority class samples in the training set are randomly undersampled under the premise of controlling the sampling ratio.

Step 2: Oversampling of positive samples. The class label variable (whether to pay or not) of the passenger data set is used as the conditional variable y of the CGAN model. For the training subset, CGAN is used to learn the distribution information of minority samples, and the training generator generates realistic minority supplementary samples to reduce the imbalance of data. Different from the traditional oversampling technology, CGAN effectively solves the problems

Figure 1. Flow chart of the proposed method



of sample overlap and overfitting, and determines the number of minority samples generated according to the number of samples and the imbalance degree in different data sets. The unpaid passenger samples with large contribution to the model are selected, and the paid passenger samples with important feature information are generated, so that the proportion of the two tends to be balanced.

3.2 Improved WOA for Feature Selection Based on Hybrid Strategy

3.2.1 Chaotic Mapping Initializes Population

The traditional whale optimization algorithm (WOA) (Mirjalili & Lewis, 2016) solves function optimization problems by initializing the population through the random number method, resulting in the initial population to be unevenly distributed in the search agent space, which greatly affects the convergence speed and the accuracy of the optimal solution of the algorithm. Chaotic motion is an irregular and ergodic nonlinear random behavior, its characteristics can improve the quality of the initial population to ensure the global search performance of the algorithm. The common chaotic maps include Logistic map, Tent map, Cubic map, Circle map and Singer map. Related research shows that the performance optimization of Tent mapping for WOA is higher than that of other mappings. Therefore, this paper uses Tent mapping to generate the initial whale population, and the mathematical expression is as follows:

$$x_{i+1} = \begin{cases} 2x_i, 0 \le x_i < 0.5 \\ 2\left(1 - x_i\right), 0.5 \le x_i \le 1 \end{cases} \tag{1}$$

For the whale population with size N and search dimension D, the tent mapping sequence $y_{ij}\left(i = 1,2,...N; j = 1,2,...,D\right)$ is generated according to Equation (1), and this sequence is mapped to the search space through Equation (2) to obtain the initial population x_{ij}:

$$x_{ij} = x_{min,j} + y_{ij}\left(x_{max,j} - x_{min,j}\right) \tag{2}$$

where y_{ij} is the chaotic variable of the jth dimension of the ith individual, $x_{min,j}$ and $x_{max,j}$ are the minimum and maximum values of the jth dimension respectively.

3.2.2 Nonlinear Convergence Factor and Adaptive Inertia Weight

In traditional WOA, parameter A is usually used to adjust the capability of the algorithm to perform global search and local development. And the convergence factor a decreases linearly with the increase of iteration times, which leads to the weak global search capability in the early stage of the algorithm and easy to fall into local optimum, in the late stage, the local development capability gradually decreases and the convergence speed Slow down. To address the shortcomings of the standard linear update strategy, a segmented nonlinear update formulation is introduced to dynamically adjust the convergence factor:

$$a = \begin{cases} 2 - (\frac{t}{T})^{\mu},\ t \le T/2 \\ 1 - \dfrac{2\left(t - \dfrac{T}{2}\right)}{\dfrac{T}{2}} + \left(\dfrac{t - \dfrac{T}{2}}{\dfrac{T}{2}}\right)^{\mu},\ t > T/2 \end{cases} \tag{3}$$

where t is the current number of iterations; T is the maximum number of iterations; μ is a nonlinear adjustment coefficient, through experiments, $\mu = 2$ is selected in this paper to balance the global search and local development capability, and enhance the convergence speed and accuracy of the algorithm.

As the traditional WOA enters the later stage of local development, due to the influence of prey on the position of whales, the individual stays near individual with better fitness in the population, thus falling into local optimum. Therefore, this paper introduces the inertia weight in the particle swarm algorithm to improve the population position update of the algorithm. The formula is as follows:

$$\omega = \left(\omega_{\max} - \omega_{\min}\right)\left\{0.8 * \left[1 - \left(\frac{t}{T}\right)^{k}\right]\right\} + \omega_{\min} \tag{4}$$

where ω_{max} and ω_{min} are the maximum and minimum inertia weights in the iterative process respectively, the control parameter k = 0.6 is selected to control the smoothness of the curve, and the inertia weights decrease with the increase as the number of iterations. At this point the position update formula becomes:

$$X\left(t+1\right) = \omega \cdot X^{*}\left(t\right) - A \cdot D, \left|A\right| < 1, p < 0.5 \tag{5}$$

$$X\left(t+1\right) = \omega \cdot X_{rand}\left(t\right) - A \cdot D_{rand}, \left|A\right| \geq 1, p < 0.5 \tag{6}$$

$$X\left(t+1\right) = D' e^{bl} \cdot \cos\left(2\pi l\right) + \left(1-\omega\right) X^{*}\left(t\right), p \geq 0.5 \tag{7}$$

3.2.3 Opposition-Based Learning Strategy

In the WOA optimization process, with the reduction of search range, the population will continue to approach the optimal solution, resulting in a decline in population diversity. Therefore, the Oppositional Based Learning (OBL) strategy is introduced to calculate the opposite solutions of individuals at the end of each iteration to improve the quality and diversity of the population, and the relative search is used instead of random search to continuously update the positions of individuals to improve the efficiency of searching the global optimal solution.

Based on the current randomly initialized population $x_{ij}\left(i = 1,2,...N; j = 1,2,...,D\right)$, the opposite population $x_{ij}^{'}$ is generated according to Equation (8):

$$x_{ij}^{'} = x_{max,j} + x_{min,j} - x_{ij} \tag{8}$$

After merging the population x_{ij} and $x_{ij}^{'}$, they are arranged in ascending order according to the fitness value, and the first *N* whale individuals with better fitness value are selected as the next iteration population.

3.2.4 Improved WOA for Feature Selection

After the class of passenger data set tends to be balanced, in order to reduce the interference of abnormal data and enhance the generalization capability of the model, this paper cross the features of some categories in passenger data sets, then combined with the original features to obtain a high-

dimensional feature set with a dimension of 1464. Feature redundancy not only affects the running speed of the algorithm, but also easily leads to overfitting. In order to filter out as few feature subsets as possible from the original high-latitude feature set, which can better represent the behavior of paid passengers, the above three kinds of search mechanisms are introduced to the traditional WOA. The improved WOA for feature selection based on hybrid strategy can be described in pseudocode as shown in Figure 2.

Considering the accuracy of the classifier and the running time of the algorithm, 20 features are finally selected as the best feature subset of passengers. Defines the behavior characteristics of known paid passengers as shown in Table 2.

3.3 Improved XGBoost Based on GHM and Genetic Algorithm

After defining the behavioral characteristics of paid passengers, in order to achieve the accurate classification of passengers by the model, this chapter mainly optimizes the XGBoost algorithm from two aspects; Firstly, GHM-C Loss is introduced to improve the loss function of XGBoost, and then the genetic algorithm is used to adjust the hyper parameters of XGBoost. Then the final recognition model of passengers' willingness to pay for seat selection is obtained.

3.3.1 GHM-C Loss Optimization

Gradient Harmonizing Mechanism(GHM)is a solution proposed by Li et al. (B. Li, Liu, & Wang, 2019) in 2019 for the imbalance between positive and negative samples, easy and hard samples in target detection tasks. Assume a data set, $p \in [0,1]$ is the probability predicted by the classification model, $p^* \in \{0,1\}$ is the true label of a class, the single cross entropy loss function is as follows:

Figure 2. Algorithm of the Improved WOA

Algorithm 1. Improved WOA for feature selection based on hybrid strategy

1. Initialize the parameters: whale population size N, dimension D, maximum iteration number T, etc. Randomly generated whale population $X_i (i = 1,2,3.....n)$
2. Initialize the whale populations by Tent Chaotic Map using Eq. (2)
3. Calculate the fitness value of each search agent to get the best search agent X^*
4. While（t < T）
5. Updates adaptive inertia weight for each search agent using Eq. (4)
6. Update convergence factor a using Eq. (3)
7. Update the other parameters: A, C, L and P
8. If（P<0.5）
9. If（|A|<1）
10. Update current position using Eq. (5)
11. Else if（|A|≥1）
12. Select a random search agent（X_r）
13. Update the position of the search agent using Eq. (6)
14. End if
15. Else if（P≥0.5）
16. Update current position using Eq. (7)
17. End if
18. End for
19. Check if any search agents are out of the search range and modify them
20. Calculate the fitness of each search agent
21. Generate opposite population using Eq. (8), update population based on fitness
22. Update X^* if there is a better solution
23. T=t+1
24. End while
25. Return X^* (Best feature subset of paid passengers)

Table 2. Behavioral characteristics of paid passengers

Feature	Meaning
pax_tax	Taxes and fees
pax_fcny	Airfare
seg_dep_time	Flight date
tkt_3y_amt	Total airfare consumption for 3 years
cabin_upgrd_cnt_y3	Number of upgrades in the last 3 years
tkt_avg_amt_y3	Average airfare spends over the last 3 years
avg_dist_cnt_y3	Average mileage per trip for the last 3 years
dist_cnt_y1	Total mileage in the past year
select_seat_cnt_y3	Number of preferred seats in recent 3 years
ssr_cnt_y3	Number of purchases of paid SSRs in recent 3 years
tkt_avg_amt_y2	Average airfare consumption in recent 2 years
cabin_f_cnt_y2	Number of first-class trips in the past 2 years
flt_bag_cnt_y3	Number of flights carrying baggage in recent 3 years
member_level	Membership level
select_seat_cnt_y2	Number of preferred seats in the past 2 years
flt_delay_time_y2	Average flight delays in the last 2 years
dist_cnt_y3	Total mileage in the past three years
seat_walkway_cnt_y3	Number of seat walkways in recent three years
mdl_mcv	passenger value
birth_date	Date of birth

$$L_{CE}\left(p,p^{*}\right)=\begin{cases}-\log\left(p\right),p^{*}=1\\-\log\left(1-p\right),p^{*}=0\end{cases} \tag{9}$$

Due to the advantages in quantity, the overall gradient contribution of simple negative samples is often much larger than that of difficult positive samples, so overwhelming the optimizer and making the training process ineffective. At the same time, the gradient modulus of particularly difficult samples (outliers) is longer than that of general samples. When the model is forced to learn to classify these samples, the classification of other samples may not be so accurate. Therefore, the gradient density function is introduced to coordinate the sample balance within a certain gradient range:

$$GD\left(g\right)=\frac{1}{l_{f}\left(g\right)}\sum_{k=1}^{N}\delta_{\in}\left(g_{k},g\right) \tag{10}$$

GHM is embedded into the loss function of the classification algorithm, and the samples that are in the larger gradient density position is regarded as simple samples to reduce the weight of outliers. The gradient density equilibrium form of the loss function is:

$$L_{GHM-C} = \sum_{i=1}^{N} \frac{L_{CE}\left(p_i, p_i^*\right)}{GD\left(g_i\right)} \tag{11}$$

According to the objective optimization function of XGBoost, this paper introduces GHM-C Loss to modify its loss function, and forms GHM-XGBoost algorithm to enhance the robustness of classification algorithm training.

3.3.2 Genetic Algorithm Optimizes Parameters

There are many parameters in the XGBoost algorithm, and the reasonable selection of hyper parameters will greatly affect the classification prediction performance of the algorithm. Therefore, this paper uses genetic algorithm to adjust the parameters of XGBoost by the average score of 5-fold cross-validation to obtains the optimal parameter combination.

In this paper, the seven parameters of learning_rate, n_estimators, max_depth, min_child_weight, gamma, sub_sample, and colsample_bytree, which have a large impact on the XGBoost model, are selected for optimization by genetic algorithm, and the other parameters are set to default values.

Combined with the improvement of XGBoost by GHM in the previous section, the main steps of using genetic algorithm to adjust XGBoost parameters are shown in Figure 3.

After optimization by genetic algorithm, the optimal combination of parameters is shown in Table 3.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Evaluation Indicator

In the general classification problem, ACC (accuracy) is used as the evaluation indicator. However, in the classification problem of such unbalanced data, ACC cannot accurately reflect the accuracy

Figure 3. Algorithm of the improved XGBoost

Algorithm 2. Improved XGBoost based on GHM and genetic algorithm

1. Input: Best feature subset of paid passengers
2. Improve XGBoost by GHM-C Loss to form GHM-XGBoost algorithm
3. Initialize algorithm parameters
4. While meet the termination condition
5. Cross-validation with XGBoost to calculate fitness
6. Select the best number of individuals for the optimal combination of algorithm parameters according to fitness
7. Genetic and mutation operations
8. Generate new algorithm parameter combinations
9. End while
10. Output: the final classification prediction model

Table 3. Optimal parameter combination

Parameter	Value	Parameter	Value
learning_rate	0.05	gamma	0.1
n_estimators	675	sub_sample	0.6
max_depth	8	colsample_bytree	0.7
min_child_weight	1		

of the model, because the imbalance of data will lead to the prediction to be biased towards the side with a large amount of data. Although the accuracy of the output is high, the accuracy is not good on a small category of samples, which is not meaningful to the actual output. Therefore, this paper selects the typical F1, G-mean and AUC as the evaluation indicators of the experiment. At present, the evaluation indicator for imbalanced data classification is obtained on the basis of confusion matrix. The confusion matrix is shown in Table 4.

According to Table 4, the following indicators can be defined:

- **F1:** The harmonic value of precision and recall, also known as F-Measure, which is closer to the two smaller ones:

$$F-score=\frac{\left(2*Precision*Recall\right)}{\left(Precision+Recall\right)} \tag{12}$$

- **G-mean:** The comprehensive index of the probability of correct classification in the positive class and the probability of correct classification in the negative class:

$$G-mean=\sqrt{\frac{TP}{TP+FN}*\frac{TN}{TN+FP}} \tag{13}$$

- **AUC:** Defined as the area under the ROC curve (ROC integral), usually greater than 0.5 and less than 1.

4.2 Result Analysis

This chapter designs two sets of comparative experiments. The first group: the experimental comparison between the improved method in stages and the previous method verifies that the improvement in each stage is feasible. The second group: The three-stage improved complete method is compared with other classical classification models to verify that the multi-stage improved ensemble method proposed in this paper has better classification effect.

4.2.1 Phased Experimental Comparison of Improved Methods

The experiment compares the original XGBoost, the improved method of one-stage data preprocessing, the improved method of two-stage feature selection and the improved complete method of three-stage algorithm. The comparison results of G-mean, F1 and AUC of each method on passenger data set are shown in Table 5.

It can be seen from Table 5 and Figure 4 that the phased improvement method proposed in this paper has a gradient increase in G-mean, F1 and AUC values, which verifies the feasibility of the phased improvement method. Among them, the fully improved G-mean and F1 are increased by 5~6% compared with those without improvement, which proves that the proposed method in this paper has

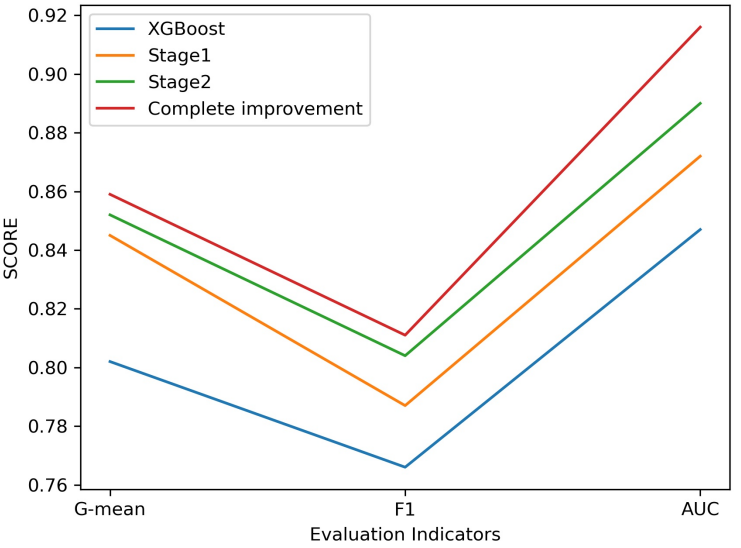
Table 4. Confusion matrix

	Actual positive category	**Actual negative category**
Predicted positive category	TP (True Positives)	FP (False Positives)
Predicted negative category	FN (False Negatives)	TN (True Negatives)

Table 5. Comparison results of phased improvements

Algorithm	G-mean	F1	AUC
XGBoost	0.802	0.766	0.847
Stage 1	0.845	0.787	0.872
Stage 2	0.852	0.804	0.890
Complete improvement	0.859	0.811	0.916

Figure 4. Phase improvement comparison results



better classification effect on the unbalanced passenger data set. At the same time, the AUC of the fully improved XGBoost is also greatly improved compared with the original XGBoost. The proposed method in this paper has better classification effect.

4.2.2 Experimental Results Based on Different Algorithms

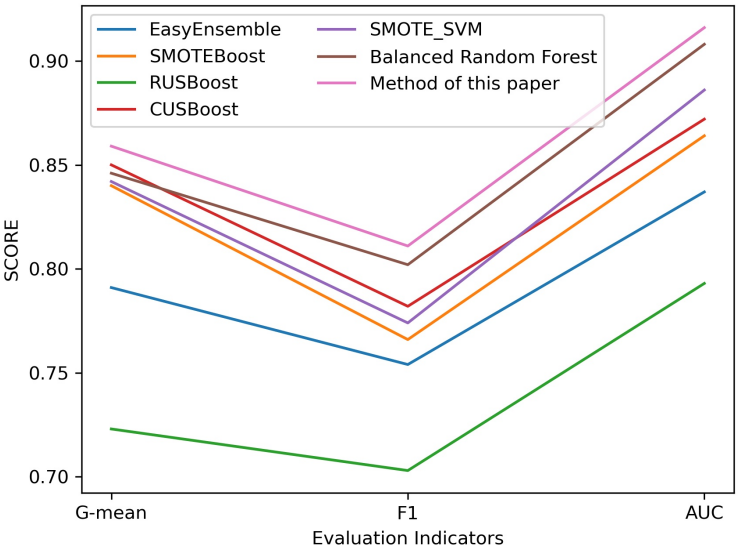
In order to prove the effectiveness of the proposed method, six classical improved algorithms EasyEnsemble, SMOTEBoost, RUSBoost, CUSBoost, SMOTE_SVM and Balanced Random Forest in the field of imbalanced data classification are selected to compare with the improved methods proposed in this paper. The comparison results of G-mean, F1 and AUC of each method on passenger data set are shown in Table 6.

It can be seen from Table 6 and Figure 5 that the improved method proposed in this paper has higher recognition rate and better classification performance than other improved classification models. In the passenger data set, the G-mean, F1 and AUC values of the proposed method are higher than those of other improved models, and the G-mean value is up to 85.9%. Moreover, the AUC value of the proposed method is significantly higher than that of other improved models, and the maximum value reaches 91.6%, indicating that the improved method proposed in this paper has good classification performance on the whole.

Table 6. Comparison results of different algorithms

Algorithm	G-mean	F1	AUC
EasyEnsemble	0.791	0.754	0.837
SMOTEBoost	0.840	0.766	0.864
RUSBoost	0.723	0.703	0.793
CUSBoost	0.850	0.782	0.872
SMOTE_SVM	0.842	0.774	0.886
Balanced Random Forest	0.846	0.802	0.908
Method of this paper	0.859	0.811	0.916

Figure 5. Comparison results of different algorithms



5. CONCLUSION

With the experience-based economy, more and more airlines are offering advance seat selection services to passengers, which will become a trend. The recognition of passengers' willingness to pay for seat selection is one of the research hotspots in the aviation business field. In order to control marketing costs and increase ancillary revenue, mining valuable customers from the large number of passenger flight records and identifying passengers with similar willingness or ability to pay is crucial for airlines to implement precision marketing.

The high-dimensional imbalance of samples is the main factor affecting the recognition of air passengers' willingness to pay for seat selection. This paper proposes a comprehensive improvement method to solve the two-category high-dimensional imbalanced data problem from three aspects: data, feature and algorithm. This method first uses the hybrid sampling of random undersampling and CGAN to balance the passenger sample at the data level; Secondly, uses feature crossover to combine features at the feature level, and then the feature subset which is more suitable for imbalanced data classification is selected by the improved whale optimization algorithm; Finally, at the algorithm

level, GHM-C loss is used to improve XGBoost, which is specifically designed for imbalanced data classification. then uses genetic algorithm to adjust the parameter and trains to obtain the final recognition model of passengers' willingness to pay for seat selection. The results of two comparative experiments prove that the improved method in this paper is feasible and has better classification effect and better performance than other improved imbalanced classification models. This study not only proves the excellent performance of XGBoost in dealing with unbalanced air passenger data, but also fully proves the superiority of multi-stage improved ensemble method in the field of passengers' willingness to pay for seat selection recognition compared with a single model, the research results of this paper provide a certain basis for the marketing of airlines' paid seat selection project.

Due to the constraints, the model in this paper was built on the basis of theoretical and historical data, so although the model achieves the initial goal of identifying potential travelers, it is not guaranteed to be fully applicable to every situation during the implementation of the project and has certain limitations. Therefore, the future research contains the following directions: (1) As the airline's database is constantly being updated, the large amount of data pre-processing work is time-consuming and requires research into more efficient algorithms for data mining. (2) As the need for explanatory models decreases for marketers, try building models using other algorithms such as neural networks, perhaps increasing the ability of the model to identify potential travelers. (3) The data used in this paper is only the consumption behavior records of travelers, and the subsequent textual information such as travelers' evaluation can be added for multi-dimensional analysis to further segment the travelers. (4) Establish a traveler willingness identification system based on the model proposed in the paper to really help decision makers develop accurate marketing strategies.

ACKNOWLEDGMENT

This work is supported by the Zhejiang Natural Science Foundation (NO. LY 20G010003) and the scientific research project of Huzhou University (NO. 2019XJWK02).

REFERENCES

Borisyak, M., Ryzhikov, A., Ustyuzhanin, A., Derkach, D., Ratnikov, F., & Mineeva, O. (2020). (1+epsilon)-class Classification: an Anomaly Detection Method for Highly Imbalanced or Incomplete Data Sets. *Journal of Machine Learning Research, 21*.

Chao, C., & Breiman, L. (2004). *Using random forest to learn imbalanced data*. University of California.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). *SMOTEBoost: Improving Prediction of the Minority Class in Boosting.* Paper presented at the European Conference on Principles of Data Mining and Knowledge Discovery. Knowledge Discovery in Databases: PKDD 2003, Berlin, Germany. doi:10.1007/978-3-540-39804-2_12

Chen, H. M., Li, T. R., Fan, X., & Luo, C. (2019). Feature selection for imbalanced data based on neighborhood rough sets. *Information Sciences*, *483*, 1–20. doi:10.1016/j.ins.2019.01.041

Chen, Z., Duan, J., Kang, L., & Qiu, G. P. (2021). A hybrid data-level ensemble to enable learning from highly imbalanced dataset. *Information Sciences*, *554*, 157–176. doi:10.1016/j.ins.2020.12.023

Cui, C., Cao, F., & Liang, G. (2020). Adaptive Undersampling Based on Density Peak Clustering. *Pattern Recognition and Artificial Intelligence*, *33*(09), 811–819. doi:10.16451/j.cnki.issn1003-6059.202009005

Dedeturk, B. K., & Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Applied Soft Computing*, *91*, 106229. Advance online publication. doi:10.1016/j.asoc.2020.106229

Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., & Bhardwaj, A. (2020). Unbalanced breast cancer data classification using novel fitness functions in genetic programming. *Expert Systems with Applications*, *140*, 112866. Advance online publication. doi:10.1016/j.eswa.2019.112866

Dongdong, L., Ziqiu, C., Bolu, W., Zhe, W., Hai, Y., & Wenli, D. (2021). Entropy-based hybrid sampling ensemble learning for imbalanced data. *International Journal of Intelligent Systems*, *36*(7), 3039–3067. doi:10.1002/int.22388

Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. *Information Processing & Management*, *57*(1), 102121. Advance online publication. doi:10.1016/j.ipm.2019.102121

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, *42*(4), 463–484. doi:10.1109/TSMCC.2011.2161285

Gao, X., Ren, B., Zhang, H., Sun, B. H., Li, J. L., Xu, J. H., He, Y., & Li, K. S. (2020). An ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling. *Expert Systems with Applications*, *160*, 113660. Advance online publication. doi:10.1016/j.eswa.2020.113660

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning.* Paper presented at the Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, Berlin, Germany. doi:10.1007/11538059_91

He, H., Yang, B., Garcia, E. A., & Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* Paper presented at the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).

Hongle, D., Yan, Z., Gang, K., Lin, Z., & Chen, Y. C. (2021). Online ensemble learning algorithm for imbalanced data stream. *Applied Soft Computing*, *107*, 107378. Advance online publication. doi:10.1016/j.asoc.2021.107378

Hosseini, E. S., & Moattar, M. H. (2019). Evolutionary feature subsets selection based on interaction information for high dimensional imbalanced data classification. *Applied Soft Computing*, *82*, 105581. Advance online publication. doi:10.1016/j.asoc.2019.105581

Jiang, Y., Liu, Y., Liu, D., & Song, H. (2020). *Applying Machine Learning to Aviation Big Data for Flight Delay Prediction.* Paper presented at the 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). doi:10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00114

Jing, X. Y., Zhang, X. Y., Zhu, X. K., Wu, F., You, X. G., Gao, Y., Shan, S., & Yang, J. Y. (2021). Multiset Feature Learning for Highly Imbalanced Data Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 139–156. doi:10.1109/TPAMI.2019.2929166 PMID:31331881

Kim, J., Kang, J., & Sohn, M. (2021). Ensemble learning-based filter-centric hybrid feature selection framework for high-dimensional imbalanced data. *Knowledge-Based Systems*, *220*, 106901. Advance online publication. doi:10.1016/j.knosys.2021.106901

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1), 273–324. doi:10.1016/S0004-3702(97)00043-X

Lee, J., & Park, K. (2021). GAN-based imbalanced data intrusion detection system. *Personal and Ubiquitous Computing*, *25*(1), 121–128. doi:10.1007/s00779-019-01332-y

Li, B., Liu, Y., & Wang, X. (2019). Gradient harmonized single-stage detector. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. doi:10.1609/aaai.v33i01.33018577

Li, L., Jiang, H., Liu, S., Zhang, J., & Xu, J. (2017). An Application of Data Miming Technology. In *Precision Marketing of Value-added Service in Aviation Industry*. Huazhong Agricultural University.

Li, X. Q., Jiang, H. K., Liu, S. W., Zhang, J. J., & Xu, J. (2021). A unified framework incorporating predictive generative denoising autoencoder and deep Coral network for rolling bearing fault diagnosis with unbalanced data. *Measurement*, *178*, 109345. Advance online publication. doi:10.1016/j.measurement.2021.109345

Liou, J. J. H., & Tzeng, G. H. (2010). A Dominance-based Rough Set Approach to customer behavior in the airline market. *Information Sciences*, *180*(11), 2230–2238. doi:10.1016/j.ins.2010.01.025

Liu, X. Y. (2006). *Exploratory Under-Sampling for Class-Imbalance Learning.* Paper presented at the Sixth International Conference on Data Mining. doi:10.1109/ICDM.2006.68

Liu, Y., & Yang, K. (2021). Credit fraud detection for extremely imbalanced data based on ensembled deep learning. *Journal of Computer Research and Development*, *58*(3), 539–547. doi:10.7544/issn1000-1239.2021.20200324

Low, R., Cheah, L., & You, L. L. (2021). Commercial Vehicle Activity Prediction With Imbalanced Class Distribution Using a Hybrid Sampling and Gradient Boosting Approach. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1401–1410. doi:10.1109/TITS.2020.2970229

Lu, L. (2018). *Research and Implementation of Loss Model of Frequent Flyers Based on Spark*. South China University of Technology.

Maliah, S., & Shani, G. (2021). Using POMDPs for learning cost sensitive decision trees. *Artificial Intelligence*, *292*, 103400. Advance online publication. doi:10.1016/j.artint.2020.103400

Mirjalili, S., & Lewis, A. (2016). The Whale Optimization Algorithm. *Advances in Engineering Software*, *95*, 51–67. doi:10.1016/j.advengsoft.2016.01.008

Nakahara, T., & Yada, K. (2011). *Extraction of Customer Potential Value Using Unpurchased Items and In-Store Movements.* Paper presented at the International Conference of Pioneering Computer Scientists, Engineers and Educators. ICPCSEE 2020: Data Science, Berlin, Germany. doi:10.1007/978-3-642-23854-3_31

Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline Over-sampling for Imbalanced Data Classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, *3*(1), 4–21. doi:10.1504/IJKESDP.2011.039875

Pan, J. Y., & Truong, D. (2021). Low cost carriers in China: Passenger segmentation, controllability, and airline selection. *Transportation*, *48*(4), 1587–1612. doi:10.1007/s11116-020-10105-z

Pang, J., Chen, K., Li, Q., Xu, Z., & Lin, D. (2021). Towards Balanced Learning for Instance Recognition. *International Journal of Computer Vision*, *129*(2), 1376–1393. doi:10.1007/s11263-021-01434-2

Pang, S. Q., & Liu, Y. J. (2011). A new hyperchaotic system from the Lu system and its control. *Journal of Computational and Applied Mathematics*, *235*(8), 2775–2789. doi:10.1016/j.cam.2010.11.029

Rayhan, F., Ahmed, S., Mahbub, A., Muhammod, R., & Farid, D. M. (2017). *CUSBoost: Cluster-Based Under-Sampling with Boosting for Imbalanced Classification.* Paper presented at the 2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions. doi:10.1109/CSITSS.2017.8447534

Razavi-Far, R., Farajzadeh-Zanajni, M., Wang, B. Y., Saif, M., & Chakrabarti, S. (2021). Imputation-Based Ensemble Techniques for Class Imbalance Learning. *IEEE Transactions on Knowledge and Data Engineering*, *33*(5), 1988–2001. doi:10.1109/TKDE.2019.2951556

Ribeiro, V. H. A., & Reynoso-Meza, G. (2020). Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. *Expert Systems with Applications*, *147*, 113232. Advance online publication. doi:10.1016/j.eswa.2020.113232

Rouncivell, A., Timmis, A. J., & Ison, S. G. (2018). Willingness to pay for preferred seat selection on UK domestic flights. *Journal of Air Transport Management*, *70*(JUL), 57–61. doi:10.1016/j.jairtraman.2018.04.018

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans, 40*(1), 185-197. 10.1109/TSMCA.2009.2029559

Shahee, S. A., & Ananthakumar, U. (2020). An effective distance based feature selection approach for imbalanced data. *Applied Intelligence*, *50*(3), 717–745. doi:10.1007/s10489-019-01543-z

Sharifai, A. G., & Zainol, Z. B. (2021). Multiple Filter-Based Rankers to Guide Hybrid Grasshopper Optimization Algorithm and Simulated Annealing for Feature Selection With High Dimensional Multi-Class Imbalanced Datasets. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 74127–74142. doi:10.1109/ACCESS.2021.3081366

Sindhu, S., & George, P. C. (2022). Analysis of Traffic Accident Features and Crash Severity Prediction. *International Journal of Cognitive Informatics and Natural Intelligence*, *15*(4), 1–18. doi:10.4018/IJCINI.20211001.oa1

Tao, Y. (2020). *Analysis Method for Customer Value of Aviation Big Data Based on LRFMC Model.* Paper presented at the International Conference of Pioneering Computer Scientists, Engineers and Educators. ICPCSEE 2020: Data Science, Singapore. doi:10.1007/978-981-15-7981-3_7

Thai-Nghe, N., Gantner, Z, & Schmidt-Thieme, L. (2010). *Cost-sensitive learning methods for imbalanced data.* Paper presented at the The 2010 International Joint Conference on Neural Networks. doi:10.1109/IJCNN.2010.5596486

Vong, C. M., & Du, J. (2020). Accurate and efficient sequential ensemble learning for highly imbalanced multi-class data. *Neural Networks*, *128*, 268–278. doi:10.1016/j.neunet.2020.05.010 PMID:32454371

Wong, M. L., Seng, K., & Wong, P. K. (2020). Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain. *Expert Systems with Applications*, *141*, 112918. Advance online publication. doi:10.1016/j.eswa.2019.112918

Wu, J., Shi, L., Lin, W. P., Tsai, S. B., Li, Y. Y., Yang, L. P., & Xu, G. S. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical Problems in Engineering*, *2020*, 1–7. doi:10.1155/2020/8884227

Wu, J., Shi, L., Yang, L. P., Niu, X. X., Li, Y. Y., Cui, X. D., Tsai, S.-B., & Zhang, Y. B. (2021). User Value Identification Based on Improved RFM Model and K-Means plus plus Algorithm for Complex Data Analysis. *Wireless Communications and Mobile Computing*, *2021*, 1–8. Advance online publication. doi:10.1155/2021/9982484

Xiong, B., Wang, G., & Deng, W. (2016). Under-Sampling Method Based on Sample Weight for Imbalanced Data. *Journal of Computer Research and Development*, *53*(11), 2613–2622. doi:10.7544/issn1000-1239.2016.20150593

Yu, M. (2019). *Research of Value-Added Service Purchase Decision of Civil Aviation Passengers Based on Customer Segmentation*. Harbin Institute of Technology.

Zhao, H. (2021). *Research on Civil Aviation Passenger Churn Model Based on Machine Learning*. Civil Aviation Flight University of China.

Zhu, Y. W., Yan, Y. T., Zhang, Y. W., & Zhang, Y. P. (2020). EHSO: Evolutionary Hybrid Sampling in overlapping scenarios for imbalanced learning. *Neurocomputing*, *417*, 333–346. doi:10.1016/j.neucom.2020.08.060

Zong, Y., & Xing, H. (2021). Customer stratification theory and value evaluation-analysis based on improved RFM model. *Journal of Intelligent & Fuzzy Systems*, *40*(3), 4155–4167. doi:10.3233/JIFS-200737

Baiyu Hong received her bachelor's degree in Computer Science and Technology from the Orient Science & Technology College of Hunan Agricultural University in June 2020. In September 2020, she started her master's degree at Huzhou University and is now a postgraduate majoring in electronic information. Her research interest includes machine learning, behavior prediction, deep learning, and artificial intelligence.

Xiaolong Ma is currently a Lecturer at the School of Business, Huzhou University, Zhejiang Province, China. He received his PhD in Management Science and Engineering at Shanghai University of Finance and Economics, China, in 2016. His interests include data mining and custom behaviour analysis.

Weining Tang holds a master's degree in computer software and theory from Northwestern University and a doctorate degree in management from Shanghai University of Science and Technology. He is a professor and master tutor. He is currently the Director of the Academic Affairs Office of Huzhou Normal University. Mainly engaged in research on innovation and entrepreneurship, supply chain management.

Zhangguo Shen is a Ph.D. candidate in the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. He received the B.E. degree in Computer Science from Zhejiang University City College, Hangzhou, China, in 2004 and the M.S. degree in Computer Science from Hangzhou Dianzi University, Hangzhou, China, in 2007. His research interests include Traffic Flow Forecasting, Machine Learning and Cloud Computing.