DR-XGBoost: An XGBoost model for field-road segmentation based on dual feature extraction and recursive feature elimination

Yuzhen Xiao[1,2†], Guozhao Mo[1,2†], Xiya Xiong[1,2†], Jiawen Pan[1,2†], Bingbing Hu[3], Caicong Wu[1,2], Weixin Zhai[1,2*]

(1. *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*;
2. *Key Laboratory of Agricultural Machinery Monitoring and Big Data Application, Ministry of Agriculture and Rural Affairs, Beijing 100083, China*;
3. *Kunlun Beidou Intelligence Technologies Co., Ltd., Beijing 102200, China*)

Abstract: Field-road segmentation is one of the key tasks in the processing of the trajectory of agricultural machinery. To improve the accuracy of the field-road segmentation, this study proposed an XGBoost model based on dual feature extraction and recursive feature elimination called DR-XGBoost. DR-XGBoost takes only a small amount of agricultural machine trajectory features as input. Firstly, the model adopted the dual feature extraction method we designed to rapidly expand the number of features and then adequately extract local trajectory features by the time window and feature extraction operator. Secondly, the model applies the recursive feature elimination algorithm to eliminate redundant features from the perspective of the model segmentation effect and thus reduce the computational consumption of model training. Thirdly, it trains XGBoost to complete the trajectory segmentation. To evaluate the effectiveness of DR-XGBoost, we conducted a series of experiments on a real trajectory dataset of agricultural machines. The model achieves a 98.2% Macro-F1 score on the dataset, which is 10.9% higher than the previous state-of-art. The proposal of DR-XGBoost fills the knowledge gap of trajectory feature extraction for agricultural machinery and provides a reasonable and effective feature selection scheme for the field-road segmentation problem.

Keywords: trajectory segmentation, feature extraction, recursive feature elimination, time window, XGBoost
DOI: 10.25165/j.ijabe.20231603.8187

Citation: Xiao Y Z, Mo G Z, Xiong X Y, Pan J W, Hu B B, Wu C C, et al. DR-XGBoost: An XGBoost model for field-road segmentation based on dual feature extraction and recursive feature elimination. Int J Agric & Biol Eng, 2023; 16(3): 169–179.

1 Introduction

In the sphere of machinery trajectory data mining, field-road segmentation is the key procedure to achieve precision agriculture and has a wide range of applications among numerous tasks. For example, effective field-road segmentation models contribute to estimating accurately the area of fields, whereby the input amount of agricultural production materials (e.g., seeds, fertilizers, etc.), the operating hours of agricultural machinery as well as the costs arising from the operation of agricultural machinery can be efficiently budgeted, making the cost of agricultural production reduce further[1-4]. Moreover, precise identification of the traveling scene of agricultural machinery can assist in adjusting the parameters related to the travel of agricultural machinery, which can reduce the consumption of fuel and the impact on the environment[5,6]. In addition, the field-road segmentation technology can extract the traveling trajectories of agricultural machines in different scenes, which combined with Internet of Things (IoTs) technology can make reasonable task assignments and timely operation evaluation for agricultural machinery[7].

The goal of field-road segmentation is to conduct semantic segmentation for the trajectory of agricultural machinery by identifying the traveling scenes of agricultural machinery. Specifically, the basic principle of field-road segmentation is to process trajectory data of agricultural machinery and identify the traveling scene of agricultural machinery at each trajectory point. Eventually, the points will be assigned the corresponding semantic labels. The trajectory refers to the sequence of spatio-temporal coordinates generated by agricultural machinery in the traveling process, the traveling scene includes operating in fields and driving on roads, and the semantic labels include the trajectory point when agricultural machines operating in fields (referred to as the field point) and the trajectory point when agricultural machinery driving on roads (referred to as the road point)[8]. Global Navigation Satellite System (GNSS) is a navigation and positioning system that provides users with 3-dimensional coordinates, as well as velocity information and time information of the object under investigation, anytime and anywhere on the Earth's surface or in near-Earth space. Related studies have shown that segmentation models based on the

Received date: 2023-02-14 **Accepted date:** 2023-05-12
Biographies: Yuzhen Xiao, Bachelor candidate, research interest: data science and big data technology, Email: xiaoyuzhen@cau.edu.cn; **Guozhao Mo**, Bachelor candidate, research interest: computer science and technology, Email: 2020309080423@cau.edu.cn; **Xiya Xiong**, Bachelor candidate, research interest: data science and big data technology, Email: xiongxiya_sia@163.com; **Jiawen Pan**, Master, research interest: computational intelligence, computer vision, Email: cau_panjiawen@cau.edu.cn; **Bingbing Hu**, Master, research interest: unmanned driving and autonomous operation of agricultural machinery, big data mining of agricultural machinery, Email: 1353232901@qq.com; **Caicong Wu**, PhD, Professor, research interest: unmanned driving and autonomous operation of agricultural machinery, big data mining of agricultural machinery, Email: wucc@cau.edu.cn
†The authors contributed equally to this work.
***Corresponding author: Weixin Zhai**, PhD, Associate Professor, research interest: big data of spatio temporal, big data mining of agricultural machinery, cartography and geographic information system. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China. Tel: +86-15810669005, Email: zhaiweixin@cau.edu.cn

traveling parameters of agricultural machinery recorded by GNSS receivers show great potential in the field-road segmentation problem, where the traveling parameters of agricultural machinery include the trajectory point location as well as speed and direction information of agricultural machinery in the process of traveling[9-11].

Despite the importance of the field-road segmentation task for agricultural machinery trajectory data mining, there are still some shortcomings in the related research so far. In the past, some studies adopted segmentation models based on the boundaries of fields to process the trajectories of agricultural machinery[12]. However, in practical applications, the boundary information of fields is difficult to obtain, which usually requires manual collection and is inefficient[13]. In addition, there are cases of misreporting and omission in the manual collection, which further affects the efficiency of the statistics and supervision for the traveling trajectory of agricultural machinery. Other studies adopted the segmentation models based on remote sensing images, i.e., using traditional image segmentation methods to segment the remote sensing images of the traveling area of agricultural machinery, but those models greatly relied on the quality of remote sensing images, and low-resolution images will significantly reduce the effectiveness of field-road segmentation[14,15].

In recent years, a few studies have proposed different machine learning models as well as deep learning models based on the traveling parameters of agricultural machinery recorded by GNSS receivers to solve the field-road segmentation problem. Specifically, these models can be classified into unsupervised machine learning models, supervised machine learning models, and supervised deep learning models. Normally, the density of agricultural machine trajectory points is high in fields owing to slower operating whereas low on roads owing to faster driving, as well as the direction is nearly parallel when agricultural machines operate in fields. Chen et al.[16] proposed an unsupervised machine learning model based on direction distribution and DBSCAN for the above characteristics of agricultural machine motion (DBSCAN+Rules), the basic idea of which is to first extract the density features of agricultural machinery trajectories using DBSCAN algorithm for initial segmentation, and secondly further correct the segmentation results by inference rules based on direction feature to achieve better segmentation effect. However, this model does not extract the motion features of agricultural machinery (e.g., acceleration, angular variation, angular velocity, angular acceleration, etc.) and classifies trajectories based on the longitude, latitude, and direction features only, which does not adequately exploit the important information in trajectory data. The density feature is reflected by spatial features (the longitude and latitude features), and the dependence on the spatial features makes the model inflexible, making the segmentation effect easily affected by the acquisition accuracy of GNSS receivers and often requiring additional correction processes for the spatial features. The model also introduces many hyperparameters with high sensitivity, whose small adjustments will dramatically affect the trajectory segmentation effect. Furthermore, the inference cannot solve the misclassification problem well due to the existence of over-correction (i.e., re-classifying a portion of correctly classified sample points as a wrong label at the same time as correcting the mis-segmented trajectory points). Poteko et al.[17] proposed a supervised machine learning model based on a decision tree (DT) which has the advantages of non-reliance on the special features of agricultural machinery, short training time, and high accuracy. Nevertheless, the model conducts inadequate and empirical-based feature extraction for trajectory data in the absence of theoretical support. Besides, the generalization performance of a single classifier applied in the model is limited, which is often inferior to that of a multi-classifier system based on integrated learning[18,19]. Chen et al.[20] proposed a supervised deep learning model based on graph convolutional neural network (GCN), which constructs a spatio-temporal graph based on temporal and spatial features of trajectory points, and then applies graph convolution to find new feature representations for the trajectory points. However, graph convolution only propagates weights between adjacent nodes, which leads to temporal and spatial scopes examined by the model being relatively limited. Besides, the features used to build the model are selected subjectively without calculation, so the model lack an objective feature selection scheme. Finally, it is usually time-consuming to train a graph convolutional neural network, which does not meet the demand for efficiency in agricultural production.

In the current trajectory data mining sphere, the theoretical researches on trajectory feature extraction are extremely sparse, and the existing trajectory feature extraction methods rely heavily on the spatial features of trajectories[21-26]. However, specifically in the sphere of agricultural trajectory data mining, the quality of GNSS receivers varies, which significantly affects the acquisition accuracy for spatial features of agricultural trajectories. Therefore, the high dependence on spatial features makes the existing trajectory feature extraction methods not suitable for direct application in the field-road segmentation problem, and the current sphere of agricultural machinery trajectory data mining lacks a complete, universal, and less equipment-requiring feature extraction method.

To address the inadequate feature extraction and empirical-based feature extraction in current studies, this paper aims to develop an XGBoost model based on dual feature extraction and recursive feature elimination (DR-XGBoost). Concretely speaking, firstly, a dual feature extraction method (DFE) is proposed in order to adequately extract the trajectory features of agricultural machinery, which is divided into two stages, motion feature extraction (MFE) and time window feature extraction (WFE). MFE rapidly extracts derived motion features based on a handful of initial motion features, thereby initially expanding the number of features. WFE further extracts time window features to capture the motion state of agricultural machinery in the local time range by time windows and feature extraction operators, thus expanding the number of features exponentially. Secondly, with the purpose of improving the efficiency of model training, we apply the recursive feature elimination algorithm based on cross-validation (RFECV) to recursively eliminate less important features from the perspective of the actual segmentation effect of the model, realizing the selection for the feature subset which makes the model effect optimal[27]. Thirdly, we input the trajectory data processed by the above feature engineering into XGBoost and construct a series of classification trees as base classifiers to segment trajectories efficiently in the form of classifier systems[28]. The main contributions of this study were as follows:

1) A dual feature extraction method with outstanding generalization performance was proposed, which effectively extracts important information from the distribution of agricultural machinery trajectory data and significantly enhances the segmentation effect of our model. To the best of our knowledge, no formalized feature extraction method for agricultural machine trajectory has been proposed before.

2) The recursive feature elimination algorithm was applied to effectively eliminate redundant features, further extracting the main

information of the distribution of agricultural machine trajectory data. The model selection causes model training less consuming in computation and further improves the segmentation efficiency of the model.

3) The highly effective feature engineering was combined with the advanced integration algorithm XGBoost to form the DR-XGBoost model which was applied to a real agricultural machinery trajectory dataset and achieved more competitive results than other field-road segmentation models.

2 Materials and methods

2.1 Datasets

This study employed the daily traveling trajectory data of agricultural machines in several Chinese provinces in the period from August to October 2019 provided by the Key Laboratory of Agricultural Machinery Operation Monitoring and Big Data Application of the Ministry of Agriculture and Rural Affairs of China as the experimental data, which contains totally 120 trajectories from Shandong Province, Henan Province, and Anhui Province, etc. The quantity of points in each trajectory is shown in Figure 1.

Each record consists of the ID of an agricultural machine, the spatial features of the agricultural machine (including the longitude and latitude features), the initial motion features of the agricultural machine (including the speed (m/s) and direction (°) features), the time feature (yyyy/mm/dd hh:mm:ss) and the artificially labeled trajectory point class (1 or 0, representing respectively the field point or road point).

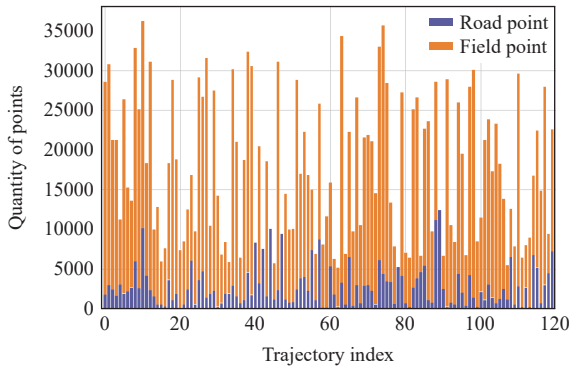


Figure 1 Quantity of points in the trajectories in the dataset

2.2 Overall framework of DR-XGBoost

As shown in Figure 2, the XGBoost model based on dual feature extraction and recursive feature elimination (DR-XGBoost) is divided into three stages: trajectory cleaning, feature engineering, and XGBoost classification. The feature engineering consists of dual feature extraction (DFE) and recursive feature elimination based on cross-validation (RFECV), where the DFE consists of motion feature extraction (MFE) and time window feature extraction (WFE).

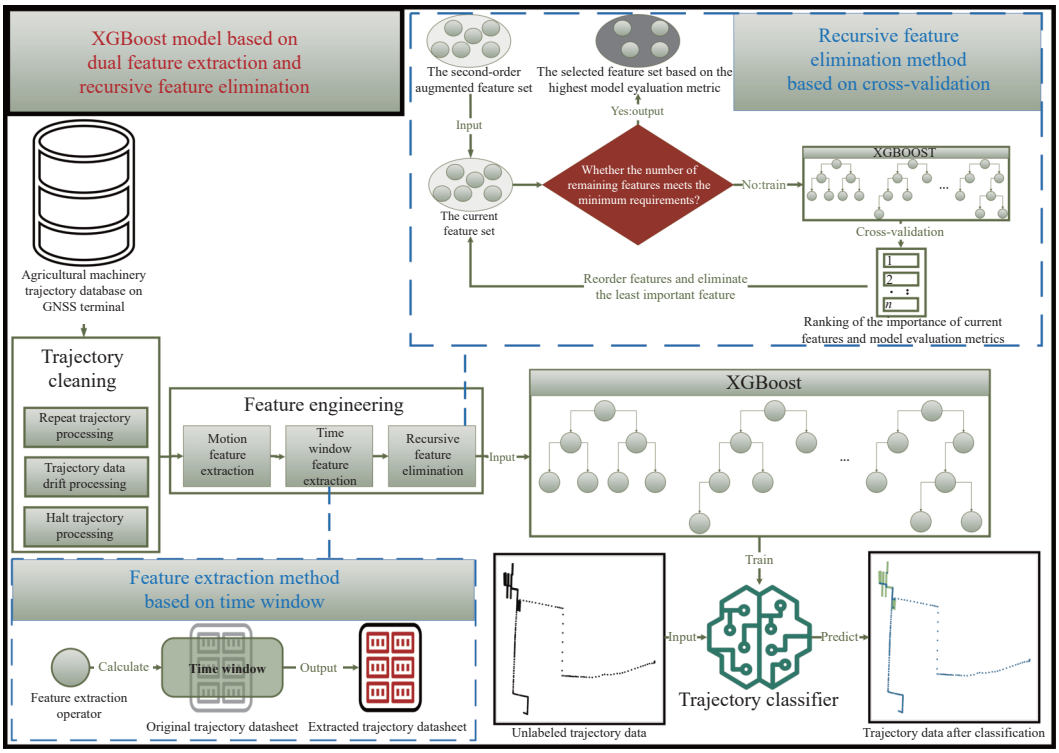


Figure 2 An overall framework of XGBoost model based on dual feature extraction and recursive feature elimination

1) Trajectory cleaning. A data cleaning process was conducted including repetitive trajectory processing, drifting trajectory processing, and stopping trajectory processing[16].

2) Motion feature extraction. To increase the number of trajectory features, MFE was conducted to deduce the derived motion features from the initial motion features thus initially expanding the feature set (presented in Section 2.3.1).

3) Time window feature extraction. In order to capture the motion state of agricultural machinery in the local time range, WFE was conducted based on time windows and feature extraction operators (presented in Section 2.3.2).

4) Recursive feature elimination. To select the most effective subset of features, the less important features added in the DFE stage are recursively eliminated in this stage (presented in Section 2.4).

5) XGBoost classification. The data processed by feature engineering is fed into XGBoost for training to obtain the trajectory classifier (presented in Section 2.5).

2.3 Dual feature extraction

2.3.1 Motion feature extraction

In order to increase flexibility, DR-XGBoost only employs the initial motion features and time features as the input of the whole model, whereby MEF is conducted to deduce derived motion features. The motion features refer to the physical quantities to describe the motion state of agricultural machinery, which can reflect the instantaneous motion state of agricultural machinery and can be further divided into the initial motion features and the derived motion features. The initial motion features include the velocity and direction features, and the derived motion features include the acceleration, angular difference, angular velocity, and angular acceleration features. The formal description of the MFE process is as follows:

A dataset with n sample data of trajectory points and 3 initial features $\mathcal{D}=\{(\boldsymbol{X}_i,t_i,y_i)|i=1,2,\ldots,n\}$ was given, where $\boldsymbol{X}_i=(v_i,\ \theta_i)$ denotes the initial motion feature of an agricultural machine at the ith trajectory point, v_i and θ_i respectively denotes the speed (m/s) and direction (steering angle, °) features of the agricultural machine at the ith trajectory point, t_i denotes the time (s) feature of the agricultural machine at the ith trajectory point, and $y_i\in\{0,\ 1\}$ (the 0 and 1 are label codes for the road point and field point respectively). Let the feature set of initial motion features be $\boldsymbol{X}^{(0)}=\{v,\ \theta\}$, and conduct MFE according to Equations (1)-(4).

$$a_i=\frac{v_i-v_{i-1}}{t_i-t_{i-1}} \tag{1}$$

$$\Delta\theta_i=\theta_i-\theta_{i-1} \tag{2}$$

$$\omega_i=\frac{\Delta\theta_i}{t_i-t_{i-1}} \tag{3}$$

$$\alpha_i=\frac{\omega_i-\omega_{i-1}}{t_i-t_{i-1}} \tag{4}$$

where, a_i, $\Delta\theta_i$, ω_i, and α_i respectively denote the acceleration (m/s^2), angular variation (°), angular velocity ((°)/s), and angular acceleration ((°)/s^2) feature of the agricultural machine at the ith trajectory point, and similarly, v_{i-1}, θ_{i-1}, and t_{i-1} respectively denote the speed and direction and time features of the agricultural machine at the (i–1)th trajectory point. The initial motion features were combined with the four derived motion features deduced from Equations (1)-(4) into the first-order augmented feature set $\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}\cup\{a,\ \Delta\theta,\ \omega,\ \alpha\}$. Due to the application of the second-order difference quotient in Equations (1)-(4), the first two trajectory points are usually discarded in MFE.

2.3.2 Time window feature extraction

It was defined that a continuous period containing a series of trajectory points as a time window. The statistics of the trajectory features within a time window can reflect the local motion state of agricultural machinery, where the statistics refer to known functions about the features of a series of trajectory points (e.g., the mean and standard deviation of local trajectory features). The traveling state of agricultural machinery can be divided into going straight and turning. In the case of going straight, agricultural machines usually operate in fields with approximately uniform motion, so the mean, median, max, and min of the acceleration, as well as the standard deviation of the velocity, is close to 0 within the corresponding time window; whereas on roads, agricultural machines often conduct accelerating and braking processes, which will result in the larger standard deviation of velocity and acceleration within the corresponding time window. Although agricultural machines may appear stationary during driving on roads (e.g., waiting for red lights or crossing pedestrians), the motion process of agricultural machines at this time includes deceleration, stop, restart, and acceleration processes, and they still have the characteristics of the larger standard deviation of velocity and acceleration within the corresponding time windows. In the case of turning, because agricultural machines operate back and forth in fields, they make frequent U-turn operations in fields, and they usually change direction slowly and evenly when making U-turns, which makes the mean of angular variation of agricultural machines in fields larger and the standard deviation of angular velocity close to zero; whereas on roads, U-turn operations of agricultural machines are not common, and there are several sharp turning operations, which makes the mean of angular variation of agricultural machines on roads smaller and the standard deviation of angular velocity larger. Based on the above analysis, the traveling characteristics of agricultural machinery on roads and in fields are significantly different both in the going straight state and the turning state, so the feature statistics difference was utilized within time windows to identify different classes of trajectory points, proposing a time-window-based feature extraction method (the corresponding process is WFE) and defining the features extracted by WFE as time window features. A formal description of WFE is as follows:

The feature extraction operator $\sigma\in F$ was introduced. F is the operator space containing all operations that can be applied to extract features. For the specific time window length $l\in\mathbb{N}_+$ (s), WFE is conducted for feature $x\in\boldsymbol{X}^{(1)}$ according to Equation (5).

$$\sigma_l(x_i)=\sigma_{t_j\in[t_i-l+1,t_i]}(x_j) \tag{5}$$

where, $\sigma_l(x_i)$ denotes the time window feature value extracted by conducting σ operation on the feature values x_j within a time window that ends at the ith trajectory point and is of length l. Denote $\sigma_l(x)$ a kind of time window feature extracted by conducting WFE on feature x using a time window with length l and the feature extraction operator σ. In this study, $\forall\sigma\in F=\{$mean, median, std, max, min$\}$, $\forall l\in L=\{200,\ 900\}$, WFE is conducted on the feature $\forall\sigma x\in\boldsymbol{X}^{(1)}$, where F denotes the actually adopted set of operators, mean, median, std, max and min respectively denote the operations of taking the mean, median, standard deviation, maximum and minimum values of the trajectory feature values within the time window, and L denotes the actually adopted set of time window lengths (the lengths are in second). The extracted time window features are added to $\boldsymbol{X}^{(1)}$ to obtain the second-order augmented feature set $\boldsymbol{X}^{(2)}=\boldsymbol{X}^{(1)}\cup\{\sigma_l(x)|\forall\sigma\in F,\forall l\in L,\forall x\in\boldsymbol{X}^{(1)}\}$.

2.4 Recursive feature elimination

DFE significantly increases the quantity of available features, but it also increases the computational consumption during model training and the possibility of feature redundancy. To eliminate redundant features in $\boldsymbol{X}^{(2)}$, DR-XGBoost conducts feature selection by applying RFECV, which recursively eliminates the less important features in combination with the feature importance provided by XGBoost[27,29]. The detailed steps are as follows:

Input: The min quantity of features to be retained p;

Output: The selected feature subset $X^{(3)}$;

Step 1 Train XGBoost using all the features in the feature set $\boldsymbol{X}^{(2)}$;

Step 2 Apply the cross-validation to evaluate the XGBoost and record the evaluation metrics;

Step 3 Calculate feature importance by the trained XGBoost model and sort the features in descending order according to the feature importance;

Step 4 For each feature subset size $s \in \{|\boldsymbol{X}^{(2)}|, |\boldsymbol{X}^{(2)}|-1, \ldots, p+1, p\}$ do

Step 5 Select the top s most important features to constitute the feature subset $\boldsymbol{X}_s^{(2)}$;

Step 6 Retrain XGBoost using the features in the feature subset $\boldsymbol{X}_s^{(2)}$;

Step 7 Evaluate the current XGBoost by cross-validation and record the current evaluation metrics;

Step 8 Recalculate feature importance by the current XGBoost model and reorder the top s most important features.

Compare the evaluation metrics of all XGBoost during the iterations and determine the optimal subset of features, which is denoted as $\boldsymbol{X}^{(3)}$;

RFECV eliminates the least important feature from the current feature set in each iteration to make the feature subset $\boldsymbol{X}_s^{(2)}$ gradually shrink until the quantity of features therein reduces to p. After completing all iterations, based on the evaluation metrics of all XGBoost, the optimal feature subset was selected as the third-order augmented feature set $\boldsymbol{X}^{(3)}$, and the feature vector of the ith trajectory point at this time is denoted as $\tilde{X}_i$.

2.5 XGBoost classification

$\boldsymbol{X}^{(3)}$ was employed as the input feature of XGBoost. The classifier system was initialized to an empty set and C rounds of iterations were conducted during training XGBoost, with a trained classification tree added to the system in each round[28]. In the cth ($c \in \{1, 2, 3, \ldots, C\}$) round of iterations, the sum of the predicted results by the first c trees was used as the probability for the ith trajectory point being a positive class in the cth round prediction $\hat{y}_i^{(c)}$ (defined in Equation (6), where $f_c(\tilde{X}_i)$ is the predicted result by the cth tree for the ith trajectory point). ω_j was denoted as the weight of the leaf node whose indexes are j and I_j as the set of all the trajectory points with predicted labels corresponding to the leaf node whose index is j. After the second-order Taylor expansion, simplification, and derivation for the objective function, the optimal weight $w_j^* = -\dfrac{G_j}{H_j + \lambda}$ was obtained, where G_j and H_j were respectively the sum of the first and second-order derivatives (of the loss function with respect to the $(c-1)$th predicted results of all trajectory points in I_j), as well as λ, is a hyperparameter.

$$\begin{cases} \hat{y}_i^{(i)} = 0 \\ \hat{y}_i^{(\theta)} = \sum\limits_{u=1}^{c} f_u(\tilde{\boldsymbol{X}}_i) = \hat{y}_i^{(a-1)} + f_0(\tilde{\boldsymbol{X}}_i) \end{cases} \tag{6}$$

Gain (defined in Equation (7)) is taken as the basis for prepruning, where γ is a hyperparameter as well as the subscripts L and R respectively denote the left and right subtrees obtained by splitting at a node, and the splitting process will be undone if Gain≤ 0 after splitting.

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_M^2}{H_k + \lambda} - \frac{(G_L + G_R)^2}{H_I + H_L + \lambda}\right] - \gamma \tag{7}$$

By the above steps, XGBoost can learn the structure and leaf node weights of each tree in each round of iterations. After completing all iterations, a classifier system consisting of C trees was obtained for trajectory segmentation.

3 Results and discussion

3.1 Validation and metrics

The Stratified K-Fold method was applied for model evaluation, dividing the training set into K subsets and ensuring the ratio of positive to negative samples in each subset was equal to that in the original dataset[30]. K rounds of evaluation were conducted, and in each round, a subset was taken from K subsets as the validation set without repetition, then merged the remaining subsets as the training set to train models and calculated metrics. After completing the K rounds, the average of the K groups of metrics was calculated as the final metric. In the experiments of this study, K=10 was taken, which was recommended by Stone et al.[31], Westerhuis et al.[32] and Neunhoeffer et al.[33], allowing to fully test the performance of models.

Five metrics were employed to evaluate the performance of models, which were precision (Pre), recall (Rec), Macro-F1 score (F1), accuracy (Acc), and training time (Time). The first four metrics were calculated based on the confusion matrix[34]. The confusion matrix consisted of true positive (TP, indicating the number of positive class samples that were correctly classified), false positive (FP, indicating the number of negative class samples that were incorrectly classified), false negative (FN, indicating the number of positive class samples that were incorrectly classified), and true negative (TN, indicating the number of negative class samples that were correctly classified). The first four metrics were calculated as Equations (8)-(11).

$$\text{Pre}_{\text{field/road}} = \frac{\text{TP}_{\text{field/road}}}{\text{TP}_{\text{field/road}} + \text{FP}_{\text{field/road}}} \tag{8}$$

$$\text{Rec}_{\text{field/road}} = \frac{\text{TP}_{\text{field/road}}}{\text{TP}_{\text{field/road}} + \text{FN}_{\text{field/road}}} \tag{9}$$

$$\text{F}_1 = \frac{\text{Pre}_{\text{field}} \times \text{Rec}_{\text{field}}}{\text{Pre}_{\text{field}} + \text{Rec}_{\text{field}}} + = \frac{\text{Pre}_{\text{road}} \times \text{Rec}_{\text{road}}}{\text{Pre}_{\text{road}} + \text{Rec}_{\text{road}}} \tag{10}$$

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

where, the subscripts field and road respectively denoted the metrics calculated with the field and road point as the positive class, whereas the metrics without subscript denoted the sum of the above two cases, and F1 and Acc were the overall metrics that combined the two classes.

3.2 Comparative experiments

DR-XGBoost was compared with several current commonly used segmentation models in various aspects. These models included GCN, DT, DBSCAN+Rules, and Random Forest (RF), where GCN was state-of-the-art agricultural machinery trajectory segmentation model, and RF was state-of-the-art city traffic trajectory segmentation model[35].

RF was trained using the dataset without feature extraction and optimized its hyperparameters by grid search[30]. The relevant settings for GCN, DBSCAN+Rules, and DT remained consistent

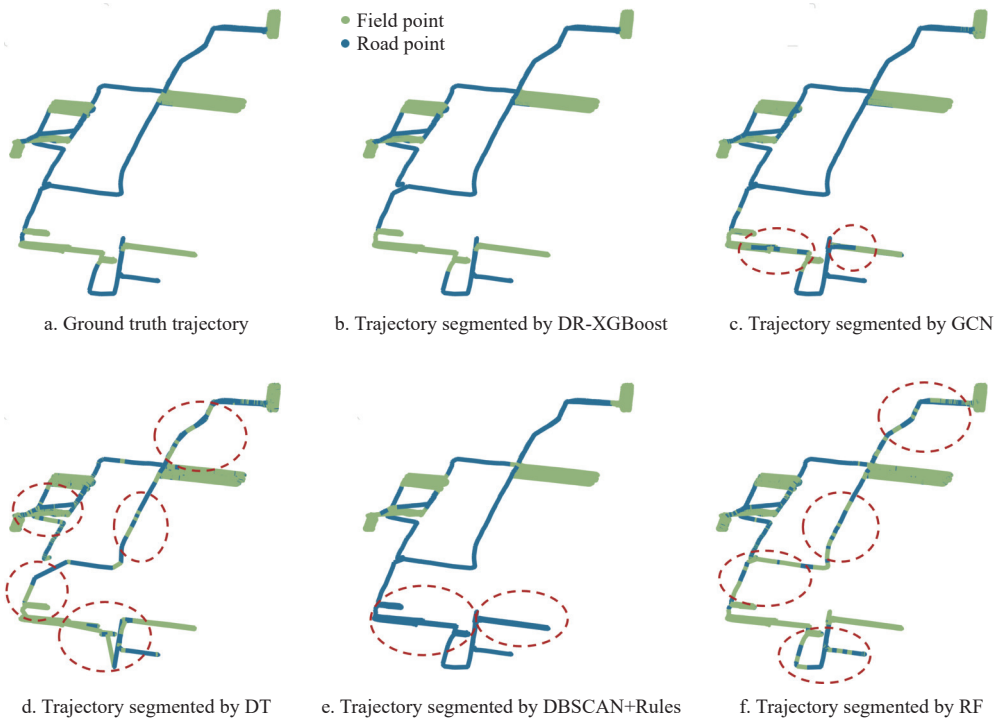
Table 1 Comparison of major segmentation models

Method	Field		Road		Accuracy	F1	Time/s
	Rec	Pre	Rec	Pre			
DR-XGBoost	99.6	99.2	96.1	97.9	99.0	**98.2**	7.38
GCN	95.4	83.3	72.1	90.4	90.3	87.9	**>3600**
DT	60.8	92.4	91.4	57.4	86.5	85.5	2.30
DBSCAN+Rules	54.1	92.3	80.3	33.5	86.3	82.7	968
RF	77.7	90.7	81.6	61.1	78.9	76.8	2.45

Note: DR-XGBoost: XGBoost model based on dual feature extraction and recursive feature elimination; GCN: Graph convolutional neural network; DT: Decision Tree; RF: Random Forest. Rec: Recall; Pre: Precision; F1: Macro-F1 score.

with the corresponding research.

The results of comparative experiments are listed in Table 1, and the rows therein were arranged in descending order according to F1. The segmentation results of a trajectory using the models are shown in Figure 3. In all metrics except Time, DR-XGBoost achieved results that outperform other models.



a. Ground truth trajectory b. Trajectory segmented by DR-XGBoost c. Trajectory segmented by GCN

d. Trajectory segmented by DT e. Trajectory segmented by DBSCAN+Rules f. Trajectory segmented by RF

Note: The area circled in red is noteworthy, where there are more misclassified trajectory points. DR-XGBoost: XGBoost model based on dual feature extraction and recursive feature elimination; GCN: Graph convolutional neural network; DT: Decision Tree; RF: Random Forest. Same below.

Figure 3 Trajectories segmented by the modes

Although the Time metric of DR-XGBoost ranked 3rd in Table 1, the difference in training time between it and RF, DT was no more than 6 s. Compared with GCN which was a deep neural network, DR-XGBoost still had a greater advantage in training efficiency. Although DR-XGBoost was moderately time-consuming, it achieved a remarkable segmentation effect.

In Figure 3a, the ground truth trajectory was given as the reference for all the segmentation models. As shown in Figure 3b, for the trajectory segmented by DR-XGBoost, the misclassified trajectory points are extremely rare for both farm points and road points.

Thanks to the support of a deep neural network, GCN outperformed the other models except for DR-XGBoost, making it effectively capture the spatio-temporal relationships between adjacent trajectory points by constructing the spatio-temporal graph and conducting the graph convolution process. As shown in Figure 3c, the trajectory segmented by GCN had fewer false field points compared to the others except for DR-XGBoost. Nevertheless, it conducted feature selection based only on prior theoretical analysis without selecting the most advantageous features from the perspective of actual segmentation effects. In addition, the spatio-temporal graph only constructed edges between adjacent trajectory points, which made the graph convolution only propagate weights between adjacent trajectory points, while DR-XGBoost employed multi-scale time windows to extract features, so the latter was easier to capture the relationship between trajectory points with larger time span than the former. In this study, DR-XGBoost actually adopted the set of time window lengths L={200, 900}. By using a time window of length 200 s, DR-XGBoost could capture the stable local motion state of agricultural machines for a short period. For some easily confused trajectory points (e.g., trajectory points generated when waiting for red lights or pedestrians on roads), DR-XGBoost could capture the change of agricultural machine motion states of a long term from time window features extracted by the time window of length 900 s to achieve correct segmentation. F1 of DR-XGBoost was 10.9% higher than GCN (listed in Table 1), and the trajectory segmented by DR-XGBoost was almost identical to the ground truth trajectory, which was a significant advantage in the segmentation effect. What was more, the difference in Time between GCN and XGBoost was multiple orders of magnitude, which made DR-XGBoost superior to GCN in terms of segmentation speed as well.

Due to the lack of reliable theoretical support, DT implemented a feature extraction process using median and standard deviation operations only in a short time range. As listed in Table 1, DT was above 90% for both Pre on field points and Rec on road points, but only around 60% for both Pre on road points and Rec on field points. As shown in Figure 3d, compared with the ground truth trajectory, there were still more false field points in the trajectory segmented by DT. The results indicated that DT tended to identify trajectory points as road points, which was a drawback brought by the inadequate feature extraction. In addition, DT used only a single classifier while DR-XGBoost used a classifier system for trajectory segmentation, so the latter had better generalization ability in most cases.

As shown in Figure 3e, the inference rule of DBSCAN+Rules enabled it to correct false field points, but excessive correction caused a mass of correctly segmented field points to be re-identify as road points, which also showed the limitation of the selected features of the model to some extent.

Benefiting from the effect of feature engineering and the excellent classification performance of XGBoost, the segmentation effect of DR-XGBoost was much better than RF. As shown in Figure 3f, the segmentation effect of RF was inferior because it had not undergone feature extraction, and compared with the ground truth trajectory, its trajectory in the field was interspersed with many false field points.

3.3 Ablation experiments

Feature engineering played a decisive role in DR-XGBoost, including MFE, WFE, and RFE. Changes in the feature set in feature engineering (including the changes of the elements and the number of elements in each order of the augmented feature set) are shown in Figure 4, and results of ablation experiments in each stage of the feature engineering are listed in Table 2, and trajectories segmented by the model in different stages of feature engineering are shown in Figure 5.

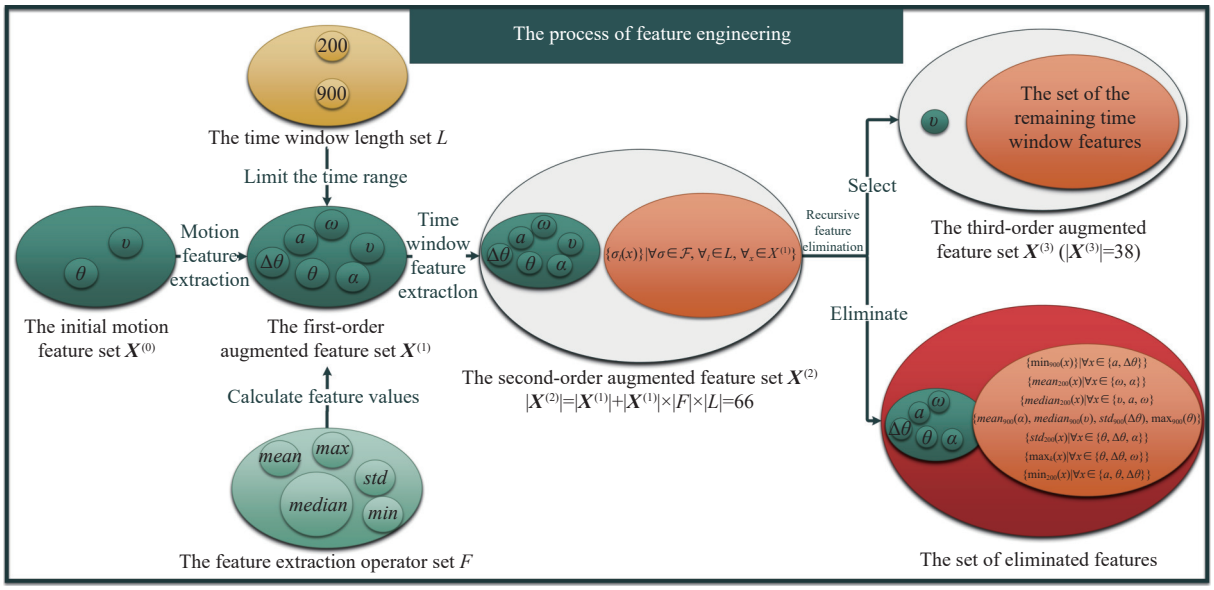


Figure 4 Changes of feature set during feature engineering

Table 2 Results of ablation experiments on each step of feature engineering

Method	Field		Road		Acc	F1
	Rec	Pre	Rec	Pre		
XGB	99.7	91.0	49.4	97.2	91.5	80.3
XGB + M	99.5	91.4	51.7	95.5	91.8	81.2
XGB + M + W	**99.7**	**98.7**	**93.0**	**98.3**	**98.6**	**97.3**
XGB + M + W + R	99.6	99.2	96.1	97.9	99.0	**98.2**

Note: The XGB represents XGBoost, the M represents motion feature extraction, the W represents time window feature extraction and the R represents recursive feature elimination based on cross-validation.

3.3.1 Motion feature extraction

The number of elements in the initial feature set $|\boldsymbol{X}^{(0)}| = 2$ (shown in Figure 4) and the handful number of unextracted features contained too little effective information to achieve satisfactory segmentation. In the ablation experiment, the ground truth trajectory is shown in Figure 5a. The unextracted initial features were input into XGBoost for segmentation, and the segmented trajectory is shown in Figure 5b. Agricultural machines usually operate in approximately rectangular fields, which makes their trajectory of them in fields usually appear continuous aggregation, but the segmented trajectory in Figure 5b obviously did not conform to the law, where the field point clusters were interspersed with numerous misclassified false road points compared with the ground truth trajectory. To preliminarily solve the problem, MFE aimed to expand the number of features and capture the instantaneous motion state of agricultural machines. XGB+M improved 0.9% on F1 compared to XGB (listed in Table 2), and a portion of the false field points was correctly segmented by adopting MFE (shown in Figure 5c), which indicated that MFE slightly improved the segmentation performance of the model. The left subplot shows two traveling trajectories of an agricultural machine, where the blue and green trajectories respectively indicate the trajectories on a road and in a field; the right subplot shows the velocity of the two trajectories in the left subplot, where the blue and green boxes respectively correspond to the blue and green trajectories.

3.3.2 Time window feature extraction

The time window feature extraction made three main contributions.

1) Further expanding the feature set;

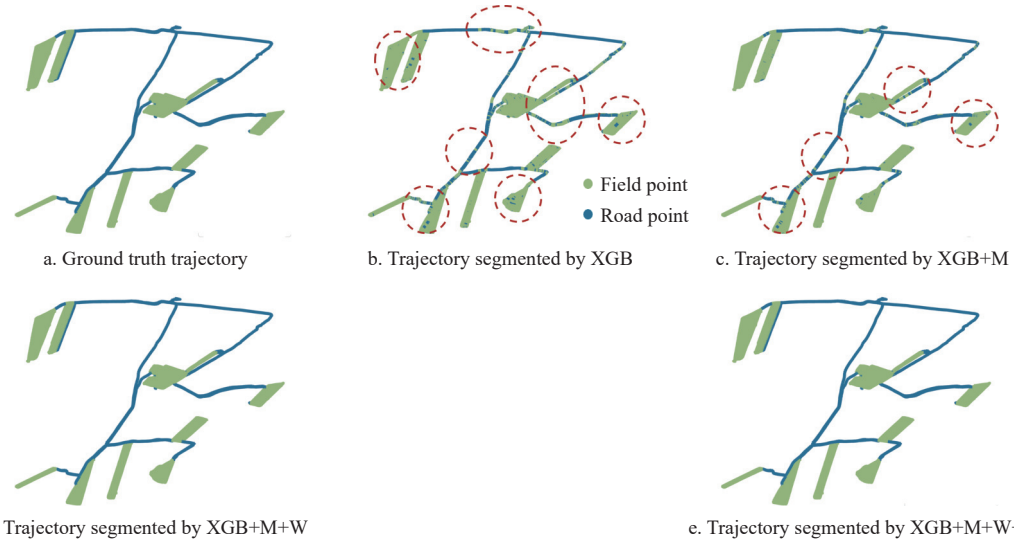
2) Capturing different types of local motion states of agricultural machines within a time window by multi-type feature extraction operators;

3) Capturing local motion states of agricultural machines at different periods by multi-scale time windows.

On the basis of MFE, WFE expanded the number of features from 6 to 66 (shown in Figure 4), exponentially increasing the available features. As listed in Table 2, XGB+M+W achieved a significant improvement in all the classification metrics, especially in terms of F1 by 16.1% compared to XGB+M, which made a favorable contribution to the final performance of the model. As shown in Figure 5d, after WFE, almost all of the false field points were correctly segmented, and the segmented trajectories were remarkably close to the ground truth trajectories.

Within any time window, different local motion states of agricultural machines were reflected as different distribution characteristics of local trajectory features (including concentration trend, dispersion degree, and range scale), which could be captured by different feature extraction operators.

The mean and median extraction operators mainly captured the concentrated trend of local trajectory features, both of which had their own advantages and shortcomings. The mean value contained



Note: *M* represents motion feature extraction; *W* represents time window feature extraction; *R* represents recursive feature elimination based on cross-validation. The area circled in red is noteworthy, where there are more misclassified trajectory points.

Figure 5 Trajectories segmented by DR-XGBoost in different stages of feature engineering

more adequate information in general because it was related to features of all trajectory points within a time window, and could represent the average level of local trajectory features. However, the mean extraction operator was susceptible to maximum and minimum feature values of local trajectory, so it was difficult for the operator to accurately express the motion state of agricultural machines when their traveling state was unstable (e.g., variable speed motion in a short period). In contrast, the median extraction operator was not affected by the extreme feature values of local trajectory, which could make up for the former deficiency, and the extracted feature values could reflect the medium level of local trajectory features, but the time complexity of the extraction process of the median extraction operator was higher than that of the mean extraction operator because the local feature values needed to be sorted before calculating the median. In most cases, using both mean and median feature extraction operators can obtain better extraction results, when there are more abnormal trajectory points in the trajectory dataset, the extraction effect of the median feature extraction operator is better; when a task required higher segmentation speed, a choice for mean feature extraction operator is more appropriate.

The standard deviation extraction operator could extract the dispersion degree of local trajectory features. As shown in the blue trajectory in the left subfigure of Figure 6, the distance difference was larger while the direction change amplitude was smaller between the consecutive trajectory points of the agricultural machine on the road, and its geographical distribution could be either dense or scattered, indicating that braking, acceleration, and sharp turning processes of the agricultural machine occurred frequently on the road, so the local features of the agricultural machine trajectory were more discrete in distribution; whereas the geographical distribution of the green trajectory points in the subfigure was more aggregated, where the distance between two consecutive points was relatively uniform, and the direction of the agricultural machine changed uniformly with larger range (the trajectory was U-shaped), indicating that the motion of the agricultural machine in the field was uniform, its change of direction was stable and it might make a U-turn. Furthermore, the box plot in the right subplot of Figure 6 respectively presented the discrete degree of the field and road trajectories, which revealed that the discrete degree of the speed of the agricultural machine trajectory on the road was obviously higher than that in the field. Therefore, the dispersion degree of local trajectory features could effectively distinguish between the two classes of trajectories,

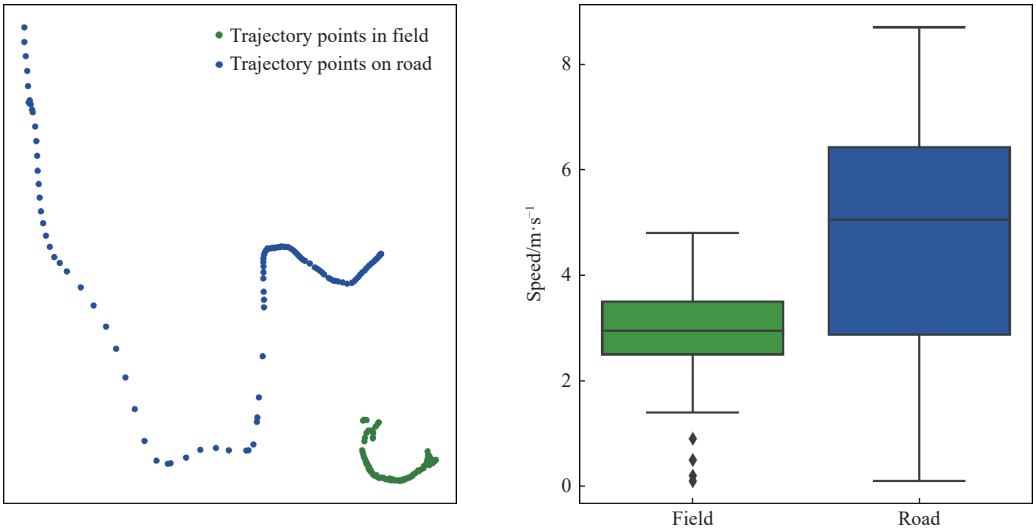


Figure 6 Comparison of two classes of trajectories

making the effect of the standard deviation extraction operator superior.

The maximum and minimum extraction operators could capture the range scale of local trajectory features. There was a noticeable difference between the speed ranges of agricultural machines in fields and on roads. As shown in the right subfigure of Figure 6, on the one hand, the operating state of the agricultural machine in fields was more stable and the operating speed was smaller, so the maximum speed within the time windows in fields was lower than that on roads; on the other hand, the agricultural machine operated continuously in fields whereas it stopped for red lights or pedestrians on roads, so the speed minimum of the agricultural machine in fields tend to be higher than that on roads during longer time windows (the speed minimum of the agricultural machines on roads was 0, see the right subfigure of Figure 6 for details). In summary, the combination of the maximum and minimum feature extraction operators helped to improve the segmentation effect of the model.

Table 3 lists the experimental results in cases of removing a single feature extraction operator, removing a type of homogeneous feature extraction operator, and retaining only a type of homogeneous feature extraction operator, where homogeneous feature extraction operators referred to feature extraction operators with the same main role and the above three cases were denoted as Case 1, Case 2, and Case 3. To begin with, the experimental results in Row 1 of Table 3 lists that the segmentation performance of DR-XGBoost could be optimal by using multi-type feature extraction operators; in case 1 (corresponding to row 2 to 6 of Table 3), removal of any operator led to a decrease in the segmentation effect of the model, with the removal of std operator causing the most severe decrease, which indicated that all the five operators contributed to the segmentation effect of DR-XGBoost, with std operator having the most significant effect; in Case 2 (corresponding to Rows 6-8 of Table 3), the removal of the range scale extraction operators (the max operator and the min operator) caused the most severe decrease, which indicated that the combination of max operator and min operator contributed more to the segmentation effect of DR-XGBoost; in Case 3 (corresponding to Row 9 to 11 of Table 3), the segmentation effect of DR-XGBoost decreased least when only the range scale extraction operator was retained, whereas it decreased most when only the discrete degree extraction operator (the std operator) was retained, which again showed the superiority of co-extraction effect of the max operator and min operator, and also illustrated the law adopted about the quantity of operators (if fewer feature extraction operators are adopted, WFE would capture less information, making the loss of segmentation performance of DR-XGBoost greater).

Table 3 Results of ablation experiments on each step of feature engineering

No.	Method	Field		Road		F1
		Rec	Pre	Rec	Pre	
1	DR-XGBoost with all the operators	99.6	99.2	96.1	97.9	**98.2**
2	(w/o) mean	99.6	99.1	95.4	97.8	98.0
3	(w/o) median	99.6	99.0	94.9	98.1	97.9
4	(w/o) min	99.6	99.0	94.9	98.1	97.9
5	(w/o) max	99.7	99.0	94.6	98.3	97.9
6	(w/o) std	99.6	99.0	94.9	97.7	**97.8**
7	(w/o) mean & median	99.6	99.1	95.1	97.7	97.9
8	(w/o) max & min	99.7	98.6	92.5	98.1	**97.2**
9	(w/o) mean & median & std	99.1	99.0	94.8	95.3	**97.0**
10	(w/o) max & min & std	99.6	97.2	85.0	97.6	94.6
11	(w/o) mean & median & max & min	99.5	97.2	85.1	97.2	**94.5**

Note: w/o: without. The mean, median, max, min, and std denote the mean, median, maximum, minimum, and standard deviation feature extraction operators, respectively.

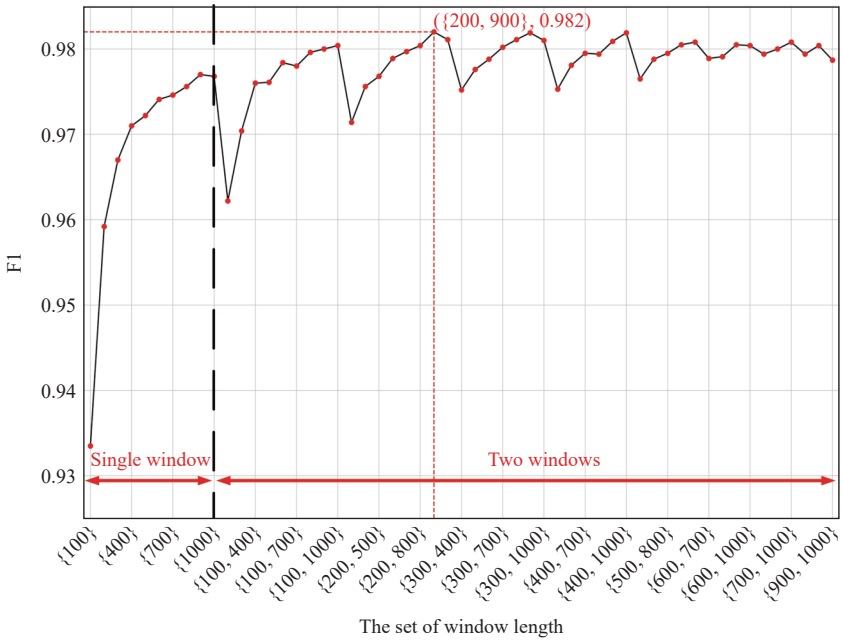
Lengths of time windows reflected the time span of feature extraction, which played a significant role in the segmentation performance of DR-XGBoost. A shorter time window is suitable for extracting short-term trajectory features, which often reflect the single-stage motion state of agricultural machines (e.g., the uniform motion state when agricultural machines are operating steadily in fields). However, for some complex cases containing multi-stage motion processes (e.g., agricultural machines leaving a field onto a road or stopping on a road to wait for a red light), it was difficult to segment trajectories accurately based on the short-term features only. To solve the above problems, DR-XGBoost introduced longer time windows to capture the multi-stage motion state of agricultural machines. The combination of multi-scale time windows effectively captured the motion states of agricultural machines in different periods, thus avoiding misclassification. The experimental results using single-scale time windows and multi-scale time windows are shown in Figure 7. The results showed that when using a single time window for feature extraction, the longer time window could lead to better segmentation effects; when further using multi-scale time windows for feature extraction, the combination of a longer time window and a shorter time window could complement each other to achieve a superior segmentation effect. Specifically, the set of time window lengths that worked optimally in this experiment was L={200, 900}, which was finally adopted.

3.3.3 Recursive feature elimination

The RFECV algorithm effectively solved the feature redundancy problem that might arise from using multi-scale time windows and multi-type feature extraction operators for WFE. Elimination of redundant features not only reduced computational consumption of the training process but also selected main features to further improve the model performance. The features eliminated by RFECV from the second-order augmented feature set $\boldsymbol{X}^{(2)}$ during the experiment were shown in Figure 4, with the quantity of remaining features $|\boldsymbol{X}^{(3)}| = 38$, which was nearly half of the number of features in $\boldsymbol{X}^{(2)}$. Furthermore, as shown in Table 2, RFECV resulted in a 0.9% improvement in F1 of the model. The above results indicated that the algorithm was able to slightly improve the segmentation performance of the model while significantly reducing the computational consumption of the model training. Finally, as shown in Figure 4, the only motion feature retained is v after RFECV. This indicated that numerous time window features were added to the feature set after WFE, which contained more important information about the trajectory, making the relative importance of motion features except for v decrease, and thus the feature redundancy problem arose.

4 Conclusions

For the field-road segmentation problem, in order to adequately extract and effectively select trajectory features of agricultural machines, this study proposed an XGBoost model based on dual feature extraction and recursive feature elimination (DR-XGBoost), where the dual feature extraction includes motion feature extraction and time window feature extraction. The motion feature extraction achieves preliminary expansion of the number of trajectory features and captures for the instantaneous motion state of agricultural machines, and the time window feature extraction achieves a



Note: The vertical axis represents the Macro-F1 score of DR-XGBoost on the experimental dataset after using corresponding time window sets, where the point marked with red horizontal and vertical dashed lines is the maximum value of the Macro-F1 score. The horizontal axis represents the sets of different time windows, and the elements in the sets represent the lengths of time windows. The left half of this figure is the results using single-scale time window sets, and the right half is the results using multi-scale time window sets.

Figure 7 Influence on segmentation effect of DR-XGBoost caused by different time window sets

significantly further expansion of the number of trajectory features and captures for different types of local motion state of agricultural machines within different periods, and the recursive feature elimination achieves selection for the optimal feature subset. Compared with other existing field-road segmentation models, DR-XGBoost presents a superior segmentation performance, and the model achieves more accurate segmentation for trajectories both in the field and on the road, which enables the Macro-F1 score of the model to be 10.9% higher than that of previous state-of-art model on the experimental dataset, showing a significant advantage.

In the future, the feature engineering for trajectory data will be continued to investigate with a view to obtaining better results on the task of field-road segmentation for agricultural machines.

Acknowledgements

This work was financially supported by the National Key Research and Development Program of China (Grant No. 2021YFB3901300), the National Precision Agriculture Application Project (Grant No. JZNYYY001), and National Innovation Training Project for University in China (Grant No. 202310019034).

[References]

[1] Bochtis D D, Sørensen C G, Busato P. Advances in agricultural machinery management: A review. Biosystems Engineering. , 2014; 126: 69-81.

[2] Molari G, Mattetti M, Lenzini N, Fiorati S. An updated methodology to analyse the idling of agricultural tractors. Biosystems Engineering, 2019; 187: 160–170.

[3] Pagare V, Nandi S, Khare D. Appraisal of optimum economic life for farm tractor: A case study. Economic Affairs, 2019; 64(1): 117–124.

[4] Sopegno A, Calvo A, Berruto R, Busato P, Bocthis D. A web mobile application for agricultural machinery cost analysis. Computers and Electronics in Agriculture, 2016; 130: 158–168.

[5] Damanauskas V, Janulevicius A, Pupinis G. Influence of extra weight and tire pressure on fuel consumption at normal tractor slippage. Journal of Agricultural Science, 2015; 7(2): 55–67.

[6] Keller T, Lamandé M, Peth S, Berli M, Delenne J Y, Baumgarten W, et al. An interdisciplinary approach towards improved understanding of soil deformation during compaction. Soil and Tillage Research, 2013; 128: 61–80.

[7] Zhang F Z, Liu R H, Ni Y D, Wang Y. Dynamic positioning accuracy test and analysis of BeiDou Satellite Navigation System. GNSS World of China, 2018; 3(1): 43–48.

[8] Wu C C, Li D, Zhang X Q, Pan J W, Quan L, Yang L L, et al. China's agricultural machinery operation big data system. Computers and Electronics in Agriculture, 2023; 205: 107594.

[9] Bochtis D D, Sørensen C G, Green O, Moshou D, Olesen J. Effect of controlled traffic on field efficiency. Biosystems Engineering, 2010; 106(1): 14–25.

[10] Grisso R D, Kocher M F, Adamchuk V I, Jasa P J, Schroeder M A. Field efficiency determination using traffic pattern indices. App.ied Engineering in Agriculture. 2004; 20(5): 563-572.

[11] Stein T, Meyer H J. Automatic machine and implement identification of an agri-cultural process using machine learning to optimize farm management information systems. In: 6th International Conference on Machine Control and Guidance, 2018; pp.19-26.

[12] Kortenbruck D, Griepentrog H W, Paraforos DS. Machine operation profiles generated from ISO 11783 communication data. Computers and Electronics in Agriculture, 2017; 140: 227–236.

[13] Kilic T, Zezza A, Carletto C, Savastano S. Missing (ness) in action: selectivity bias in GPS-based land area measurements. World Development, 2017; 92: 143–157.

[14] Rydberg A, Borgefors G. Integrated method for boundary delineation of agricultural fields in multispectral satellite images. IEEE Transactions on Geoscience and Remote Sensing. 2001; 39(11): 2514-2520.

[15] Yan L, Roy D P. Automated crop field extraction from multi-temporal Web Enabled Landsat Data. Remote Sensing of Environment, 2014; 144: 42–64.

[16] Chen Y, Zhang X Q, Wu C C, Li G Y. Field-road trajectory segmentation for agricultural machinery based on direction distribution. Computers and Electronics in Agriculture. 2021;186: 106180.

[17] Poteko J, Eder D, Noack PO. Identifying operation modes of agricultural vehicles based on GNSS measurements. Computers and Electronics in Agriculture, 2021; 185: 106105.

[18] Schapire R E. The strength of weak learnability. In: 30th Annual Symposium on Foundations of Computer Science, Research Triangle Park, 1989; pp.28-33.

[19] Valiant L G. A theory of the learnable. Communications of the ACM, 1984; 27(11): 1134–1142.

[20] Chen Y, Li G Y, Zhang X Q, Jia J P, Zhou K, Wu C C. Identifying field and road modes of agricultural Machinery based on GNSS Recordings: A graph convolutional neural network approach. Computers and Electronics in Agriculture, 2022; 198: 107082.
[21] Feng Z N, Zhu Y M. A survey on trajectory data mining: Techniques and applications. IEEE Access, 2016; 4: 2056–2067.
[22] Lee J G, Han J, Li X, Gonzalez H. TraClass: Trajectory classification using hierarchical region-based and trajectory-based clustering. Proceedings of the VLDB Endowment, 2008; 1(1): 1081–1094.
[23] Mazimpaka J D, Timpf S. Trajectory data mining: A review of methods and applications. Journal of Spatial Information Science, 2016; 13: 61–99.
[24] Wang D, Miwa T, Morikawa T. Big trajectory data mining: A survey of methods, applications, and services. Sensors, 2020; 20(16): 4571.
[25] Wang S, Bao Z F, Culpepper J S, Cong G. A survey on trajectory data management, analytics, and learning. ACM Computing Surveys, 2021; 54(2): 39.
[26] Zheng Y. Trajectory data mining: An overview. ACM Transactions on Intelligent Systems and Technology (TIST), 2015; 6(3): 29.
[27] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using Supp.rt Vector Machines, Machine Learning, 2002; 46: 389-422.
[28] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; pp.785-794.
[29] Duan K B, Rajapakse J C, Wang H Y, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Transactions on NanoBioscience, 2005; 4(3): 228–234.
[30] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 2011; 12: 2825–2830.
[31] Stone M. Cross-validation: A review. Series Statistics, 1978; 9(1): 127–139.
[32] Westerhuis J A, Hoefsloot H C, Smit S, Vis D J, Smilde A K, van Velzen E J, et al. Assessment of PLSDA cross validation. Metabolomics, 2008; 4: 81–89.
[33] Neunhoeffer M, Sternberg S. How cross-validation can go wrong and what to do about it. Political Analysis, 2019; 27(1): 101–106.
[34] Tharwat A. Classification assessment methods. Applied Computing and Informatics, 2021; 17(1): 168–192.
[35] Dabiri S, Markovic N, Heaslip K, Reddy C K. Á deep convolutional neural network based approach for vehicle classification using large-scale GPS trajectory data. Transportation Research Part C:Emerging Technologies, 2020; 116: 102644.