Proceedings of CONF-MSS 2025 Symposium: Automation and Smart Technologies in Petroleum Engineering
DOI: 10.54254/2755-2721/2025.KL22356

A Study of Robot Ground Classification Based on SMOTE Oversampling Technique and XGBoost

Zongbo Yu

China University of Mining & Technology, Beijing, China
783984598@qq.com

Abstract: Accurate identification of ground types during robot travelling is crucial for improving the robot's navigation stability and decision-making ability. To this end, this paper proposes a robot ground type classification method based on SMOTE oversampling technique and XGBoost, specifically, we firstly balanced the unbalanced dataset by SMOTE technique, and thereafter fed the processed data into XGBoost model for classification. We conducted extensive comparative experiments comparing common machine learning algorithms and found that the 91% accuracy of our algorithm achieved the best results, which proves the advancement and superiority of our proposed method and provides effective technical support for autonomous robot navigation and environment sensing.

Keywords: Machine Learning, XGBoost, SMOTE oversampling technique

1. Introduction

With the industrial development and the improvement of robots' autonomous mobility, their applications in industrial inspection, agricultural automation, medical aid and other fields are becoming more and more widespread [1-3]. However, ground classification under complex terrain is still one of the core challenges in robot navigation, which directly affects its travelling stability and path planning [4]. In order to improve the accuracy of ground type classification and route planning, it is crucial to propose an algorithm that effectively classifies robot ground types.

In past studies, ground type classification methods can be mainly classified into traditional machine learning methods based on mathematics and deep learning methods based on neural networks [5]. Huang et al. [6] proposed a classification method based on Support Vector Machines (SVMs) and Binary Tree Clustering to classify the grasping patterns of robotic multi-fingered hands, Li et al. [7] classified China's Southern hilly land terrain and found that the classification accuracy of random forest is better than SVM and KNN algorithms; Yang et al. [8] compared nine machine learning methods and selected XGBoost integrated learning algorithm with the highest eigenvalue score to construct a strong convective weather classification and identification model to classify and identify different weather, and examined the good prediction performance; Cheng [9] based on a Depth-separated convolutional neural network to extract flower features, using a miniature embedded device as the robot's "eyes", to achieve the robot's automated classification technology for flowers; Xu et al. [10] used a deep recurrent neural network in the long and short-term memory model for collaborative robot dynamics model to compensate for the error, and the compensated collaborative robot dynamics model has good performance for the actual moment. Robot dynamics model has a better prediction of actual moments. Liu [11] compared the effect of both on large sample and small

DOI: 10.54254/2755-2721/2025.KL22356

sample image scene classification using traditional machine learning methods and deep learning methods respectively, and found that the traditional machine learning has a better solution on small sample dataset, while the deep learning framework has a higher recognition accuracy on large sample dataset.

However, in practical applications, deep machine learning is not suitable for real-time applications due to high arithmetic consumption and high training costs. The classification of traditional maths-based machine learning is not accurate enough and features are not extracted sufficiently. In addition we found that the samples are not always balanced in the dataset [12] and most of the previous work does not consider this in detail, we need synthetic minority oversampling technique (SMOTE) to bridge the gap between the minority and the majority of samples [13].

Based on the above analysis, this paper proposes a robot ground type classification method based on SMOTE oversampling technique and XGBoost. Specifically, we first 2inimize the unbalanced dataset with the SMOTE method, after which we actively extract common statistical features and then train the classification with the XGBoost algorithm, which uses the characteristics of the gradient boosting tree to enable efficient and accurate classification on different types of ground data, and the computational cost is much lower than that of the neural network model, which is efficient and practical.

2. Methodology

XGBoost (eXtreme Gradient Boosting) is an efficient implementation of Gradient Boosting, proposed by Chen and Guestrin in 2016 [14] , which is an integrated learning method that gradually improves the model by constructing a set of weak learners (usually decision trees) mainly to improve the model performance to achieve the final strong classifier, using gradient-based optimisation to minimise the loss function, and combining with a regularisation strategy to improve the generalisation ability of the model. Compared with the traditional Gradient Boosting Tree (GBDT), XGBoost provides significant improvements in performance and computational efficiency.

2.1. XGBoost

XGBoost (eXtreme Gradient Boosting) is an efficient implementation of Gradient Boosting, proposed by Chen and Guestrin in 2016 [14] , which is an integrated learning method that gradually improves the model by constructing a set of weak learners (usually decision trees) mainly to improve the model performance to achieve the final strong classifier, using gradient-based optimisation to minimise the loss function, and combining with a regularisation strategy to improve the generalisation ability of the model. Compared with the traditional Gradient Boosting Tree (GBDT), XGBoost provides significant improvements in performance and computational efficiency.

2.1.1. Design Ideas

XGBoost is constructed based on an additive model and a forward-stepping algorithm, which aims to minimise the following objective function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

Among them:

- $l(y_i, \hat{y}_i)$is the prediction error (e.g. log loss) of the training sample$(\mathrm{x_i}, \mathrm{y_i})$.
- $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$is the predicted value for the ith sample and$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$ is the prediction for the kth tree.

DOI: 10.54254/2755-2721/2025.KL22356

- $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ is the regularisation term of the model complexity, where T is the number of leaf nodes of the tree andw_j is the weight of the jth leaf node.

2.1.2. Objective function optimisation

In the tth iteration, XGBoost fits a new treef_t to minimise the following second-order expansion of the objective function on top of the already constructed$t-1$ tree:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \tag{2}$$

Among them:

- $g_i = \frac{\partial l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \widehat{y_i^{(t-1)}}}$is the first order derivative (gradient).
- $h_i = \frac{\partial^2 l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \widehat{y_i^{(t-1)2}}}$is the second order derivative (diagonal element of the Hessian matrix).

By introducing the second-order derivative information, XGBoost achieves the approximate optimisation of the objective function, which significantly improves the optimisation efficiency and model performance.

2.1.3. Selection of the splitting point

XGBoost uses a greedy algorithm to select split points with the goal of maximising the objective function gain (Gain). Gain is defined as follows:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \tag{3}$$

Among them:

- G_L,H_L and,$G_R H_R$ are the first and second order gradient sums of the left and right child nodes respectively.
- λ and γ are regularisation parameters to control the complexity of the model.

The gain calculation process ensures that each split maximises the optimisation of the objective function while avoiding overfitting.

2.1.4. Regularisation

XGBoost introduces two regularisation parameters λ and γ in the objective function:

- λ controls the L2 regularisation of the leaf node weights and is used to suppress overfitting due to too large weights.
- γ controls the complexity of splitting, and the splitting operation is performed only if the splitting gain exceeds γ.

This regularisation mechanism enhances the generalisation ability of the model and enables it to achieve good performance on both training and test data.

DOI: 10.54254/2755-2721/2025.KL22356

2.2. SMOTE oversampling technique

SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling technique for dealing with class imbalance problem [15], which balances the distribution of classes in a dataset by synthesising new samples of the minority class, the algorithm proceeds as follows.

2.2.1. Selecting a small sample of classes and calculating neighbours

First, the algorithm selects a target sample from the minority samples in the dataset: assuming that the set of minority samples in the dataset is$S_{minority}$, for each target minority sample$x_i \in S_{minority}$, the SMOTE algorithm calculates the distance of that sample from all other minority samples:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_i^k - x_j^k)^2} \tag{4}$$

Where,x_i^k andx_j^k are the values of samplesx_i andx_j on the kth feature and n is the total number of features.

2.2.2. Generating new samples

For each target samplex_i , a sample$x_{i'}$ is randomly selected from its k nearest neighbour samples and a new samplex_{new} is generated by linear interpolation:

$$x_{new} = x_i + \lambda \times (x_{i'} - x_i) \tag{5}$$

Among them:

- λ is a randomly chosen value for$\lambda \in [0,1]$.
- x_i is the target sample.
- $x_{i'}$ is a sample of selected neighbours.

This formula indicates that the new samplex_{new} is located between the target samplex_i and its neighbour$x_{i'}$ and the interpolation factorλ controls where the new sample is generated.

2.2.3. Repeatedly generating samples and synthesising the training set

The SMOTE algorithm keeps repeating the previous step until it generates a sufficient number of minority class samples in order to reach a predetermined balancing goal. By generating new samples several times, SMOTE takes the newly generated synthetic samples together with the original minority class samples to form a new training set which will have a more balanced class distribution.

3. Experiments

3.1. Experimental hardware and software environment

The experiments were conducted in a GPU environment of NVIDIA GeForce RTX 4060, the programming language was python and the deep learning framework was pytorch version 2.5.1.

DOI: 10.54254/2755-2721/2025.KL22356

3.2. Data processing

3.2.1. Introduction to the dataset

The dataset of this experiment contains 3810 groups of signals collected from the robot, each group of signals has 10 channels and each channel has 128 sampling points, i.e. each group of signals is a 128*10 signal matrix. The output is the category of the ground to which each group of signals belongs, and the nine categories and their corresponding quantities are shown in Fig. 1.

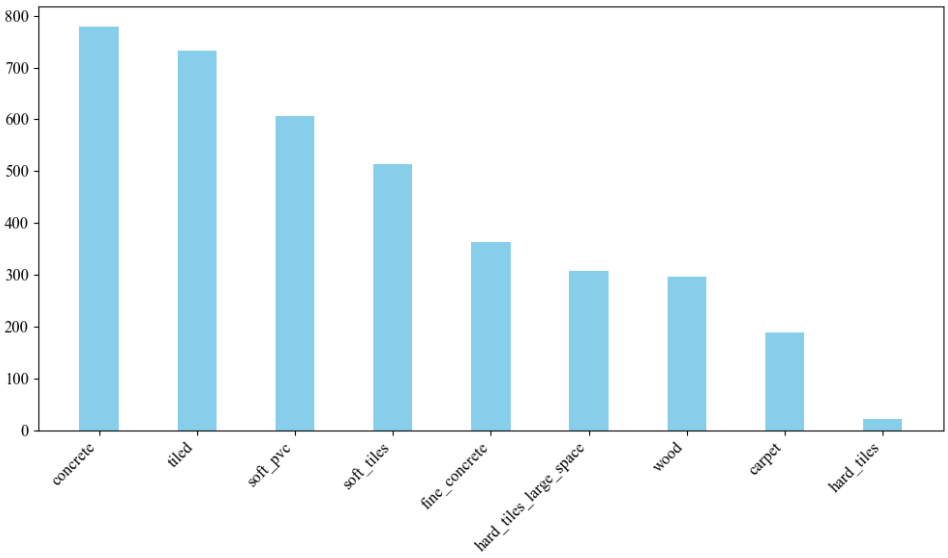

Figure 1: Distribution of number of sample categories

3.2.2. SMOTE strategy data balancing

From Fig. 1, it can be found that the number of samples in each category is unevenly distributed. To solve this problem, we adopt the data balancing (SMOTE) strategy [16] to balance the number of samples in the training set, generating enough samples from few categories to make the number of samples in each category equal, Figure 2 shows the proportion of each category in the dataset before and after balancing.

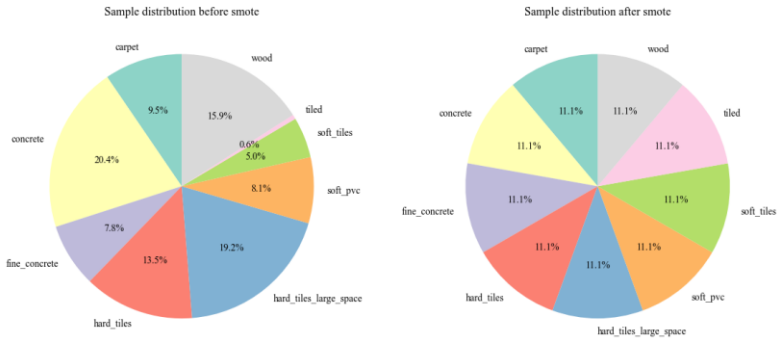

Figure 2: Comparison of sample category distributions.

DOI: 10.54254/2755-2721/2025.KL22356

3.2.3. Feature extraction

Feature extraction is the core part of data preprocessing. In order to compress matrix information and save training cost, we extract multiple statistical features from each group of signals, such as mean, standard deviation, maximum, minimum, skewness and kurtosis. These features can effectively describe the trend, volatility, distribution characteristics and peaks of the signal, which are the core characteristics of the signal, and the extraction of these features is conducive to the learning of the subsequent model.

Finally, we divided the dataset into 80% training set and 20% testing set.

3.3. Hyperparameter selection

The hyperparameters of the XGBoost classifier are set as follows:

- Learning_rate: set to 0.1 to control the weight of each tree to avoid the model converging too quickly.
- Maximum tree depth (max_depth): set to 6 to limit the depth of a single tree to prevent overfitting.
- Number of base learners (n_estimators): set to 100, indicating that up to 100 trees are constructed.
- Evaluation metric (eval_metric): set to mlogloss, the multicategorical log-loss, which measures the predictive performance of the model.
- Random seed (random_state): set to 42, used to ensure the reproducibility of the experimental results.

3.4. Experimental indicators

In order to effectively assess the performance of the model, several experimental metrics were used in this experiment to judge the classification effect, specifically, the stipulation:

TP (True Positives): True cases, which are predicted to be positive and actually are;

FP (False Positives): False Positive cases, where the prediction is positive but the actual case is negative;

FN (False Negatives): False Negatives, which are predicted to be negative but are actually positive;

TN (True Negatives): True Negatives, which are predicted to be negative and actually are.

Categorical accuracy indicates the ratio of the number of correct predictions to the total number of positive and negative examples:

$$\text{accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{6}$$

The precision rate is the proportion of correct predictions in a sample of positive cases where the prediction is judged on the basis of the outcome of the prediction:

$$\text{precision} = \frac{TP}{TP+FP} \tag{7}$$

Recall is the proportion of correctly predicted positive cases out of the total sample of **actual** positive cases, judged on the basis of **actual samples**:

$$\text{recall} = \frac{TP}{TP+FN} \tag{8}$$

The F1 score is a metric that neutralises precision and recall. Where P is the precision rate and R is the recall rate:

$$\text{F1} = \frac{2PR}{P+R} \tag{9}$$

DOI: 10.54254/2755-2721/2025.KL22356

3.5. Comparative experiments

As shown in Table 1, we compare our method with svm and convolutional neural network models , and find that our method outperforms the remaining two methods in all common metrics, proving the superiority of our proposed algorithm.

Table 1: Comparison of results for XGBoost, SVM and Convolutional Neural Network.

	Classification accuracy	Accuracy	Recall rate	F1score
XGBoost+SMOTE	0.91	0.91	0.91	0.91
SVM	0.59	0.58	0.59	0.58
CNN	0.77	0.78	0.77	0.77

4. Conclusion

In this paper, we use the public dataset provided by the CareerCon 2019 competition in the kaggle website as the experimental dataset, and propose a robot ground type classification method based on SMOTE oversampling technique and XGBoost. The experimental results show that the recognition ability of the model is stable regardless of minority class ground types or majority class ground types, and the accuracy and F1 score in terms of the overall classification performance are 91% and 0.91, respectively, which are higher than 59% and 0.58 of SVM and 77% and 0.77 of CNN.In addition, it is also found that reasonable feature engineering and data balancing processing have a significant influence. This study can provide reliable support for autonomous robot navigation, path planning, and environment sensing, and can be extended to other classification tasks with data imbalance. Future research can further optimise the feature extraction method and combine it with other deep learning models to improve the ground recognition capability in complex environments.

References

[1] *Arents J, Greitans M. Smart industrial robot control trends, challenges and opportunities within manufacturing[J]. Applied Sciences, 2022, 12(2): 937.*

[2] *Ning W, Yuxiao H A N, Yaxuan W, et al. Research progress of agricultural robot full coverage operation planning [J]. Nongye Jixie Xuebao/Transactions of the Chinese Society of Agricultural Machinery, 2022.*

[3] *Guo Y, Chen W, Zhao J, et al. Medical robotics: opportunities in China[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2022, 5(1): 361-383.*

[4] *Karur K, Sharma N, Dharmatti C, et al. A survey of path planning algorithms for mobile robots[J]. Vehicles, 2021, 3(3): 448-468.*

[5] *Li H. Machine learning methods [M]. Tsinghua University Press, 2022.*

[6] *Yuanjie Huang,Congqing Wang. Support vector machine-based pre-grasping pattern classification for robotic multi-fingered hands[J]. Mechanical Engineering and Automation, 2006,(04):94-96.*

[7] *Hengkai LI, Lijuan WANG, Songsong XIAO. Random forest classification of land use in southern hilly mountains based on multi-source data[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(7).*

[8] *He YANG, Hongbo MA, Weinan SUN, et al. Classification and identification of strong convective weather in Jilin Province based on XGBoost algorithm[J]. Meteorological Disaster Defense,2023,30(02):28-33.*

[9] *Cheng T. Robotic flower sorting system based on depth-separated convolutional neural network[D]. Hunan University of Technology,2019.*

DOI: 10.54254/2755-2721/2025.KL22356

[10] XU Z, ZHANG G, WANG H, et al. Error compensation of collaborative robot dynamics based on deep recurrent neural network[J]. Chinese Journal of Engineering, 2021, 43(7): 995-1002.

[11] Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning[J]. Pattern recognition letters, 2021, 141: 61-67.

[12] Siwei XU, Ming ZHOU, Rui ZOU, et al. A study on the effect of sampling ratio on classification results in unbalanced datasets[J]. Intelligent Computers and Applications,2024,14(09):111-117.DOI:10.20169/j.issn.2095-2163.240917.

[13] Aihua Li, Wanxin Li, Sifan Chen, et al. Research on SMOTE-BO- XGBoost integrated credit scoring model for unbalanced data[J/OL] China Management Science,1-10[2025-2-13]Chen, Tianqi and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): n. pag.

[14] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[15] Chawla, N., K. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." ArXiv abs/1106.1813 (2002): n. pag.

[16] ZHANG T, DING L. A new resampling method based on SMOTE for imbalanced data set [J][J]. Computer Applications and Software, 2021, 38(9): 273-279.