ISSN: 2688-8645

Attitude Detection Classification based on Feature Selection and Integrated Machine Learning Model

Yonglin Zou, Xuefen Liu, Ziqi Liao, Zhi Li*

School of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang, China

*1683680099@qq.com

Abstract

This paper classifies 19 human actions based on a dataset of 1.2 million human actions acquired from sensors. The Xgboost classification model is established, and the MIV algorithm is used as an index to evaluate the importance of each variable to the dependent variable. The MIV values of all features are sorted according to the absolute value of each variable, and finally the top 10 groups of features are selected as the features of the reduced data set, and fine-tuned by grid search. Select the simplified data set, get the maximum roc_auc through continuous testing, and get the optimal model. The model has a recall of 1 and a precision, F1-score, and AUC of 0.99. Then, in order to make the model have a good generalization ability under the limited data set, a feasible method is designed to evaluate the generalization ability of the model. using the SMOTE-Tomek integrated sampling method to calculate k-nearest neighbor samples for each minority class sample, select the class samples whose neighbor similarity coefficients meet the requirements. Randomly generate new samples according to the adjacency relationship between the linearly interpolated sample and its neighbor samples. According to the original data training set, generate similar data samples and put them into the model trained with the original data for prediction evaluation. Finally, the classification accuracy rate of the generated samples is obtained, the precision rate is 0.98, and the recall rate is 0.99. The F1 value is 0.98, and the Roc_AUC value is 0.98, which proves that the evaluation model has good generalization ability.

Keywords

MIV; Xgboost; Integrated; Grid Search; Comprehensive Sampling.

1. Introduction

Human behavior understanding is an important context for recognizing and monitoring our daily lives, and with the help of wearable activity recognition systems, our quality of life can be improved in many ways. In [1], the researchers collected a data set of 19 different actions of 1.2 million human bodies through the micro-inertial sensor and magnetometer on the wearable device. This paper attempts to deduce the activity state of the human body from this dataset, that is, to accurately classify the data collected during various activities.

Each of the 19 activities was performed for five minutes by eight subjects, four females and four males, aged between 20 and 30 years. The total signal duration was 5 min per activity for each subject. Subjects were asked to perform the activity in their own style and were not limited to how the activity was performed. For this reason, there are inter-subject differences in the speed and magnitude of some activities.

The sensor unit [2] was calibrated to collect data at a sampling frequency of 25 Hz. The 5-min signal was divided into 5-second segments, resulting in 480 (=60×8) signal segments for each activity.

The 19 activities include:

ISSN: 2688-8645

Sitting posture (A1); Stand (A2); Supine (A3); Right side (A4); Upstairs (A5); Descending stairs (A6); Standing still in the elevator (A7); Walk in the elevator (A8); Walking in the parking lot (A9);

Walking on a treadmill at 4 km/h in a flat position and a 15-degree tilt position (A10); Walking on a treadmill at 4 km/h in a 15-degree tilt position (A11); Run on a treadmill at 8 km/h (A12). Exercise on a stepper (A13); Work out on a cross-trainer (A14). Riding on a horizontal exercise bike (A15);

Riding on an exercise bike in a vertical position (A16); Boating (A17); Jump (A18); Playing basketball (A19).

To solve this classification problem, we divide it into 2 tasks:

Task 1: In order to classify 19 human behaviors based on the data from these wearable sensors, this paper trains the Xgboost model by building the Xgboost classification model and dividing the dataset into 70% training set and 30% test set. The Xgboost model for human behavior classification is obtained.

Task 2: Due to the high cost of data, we need to use a limited data set to make the model have good generalization ability, so we need to specifically study and evaluate this problem, and design a feasible method to evaluate the generalization ability of the model.

2. Models Introduction

Since the wearable sensor data comes from five units on the torso (T), right arm (RA), left arm (LA), right leg (RL), left leg (LL), each with nine sensors (x, y, z accelerometers, x, y, and z gyroscopes, x, y, and z magnetometers), Therefore, each row contains all the data collected from 45 sensors at a specific sampling time, and we need to reduce the dimension of these 45-feature data. Therefore, this paper uses MIV algorithm to extract the importance of each feature for human activity classification, and filters the features to achieve data dimension reduction.

The MIV algorithm [3] is an indicator used to determine the influence of the input features on the output category of activity, with the symbol representing the direction of correlation and the absolute value representing the relative importance of the influence. The calculation process: After the model training is terminated, each independent variable feature in the training sample P is added and subtracted by 10% from its original value to form two new training samples P1 and P2, and P1 and P2 are classified as new data sets using the established Xgboost model, and the classification results of the two data sets A1 and A2 are obtained. Finally, the MIV is averaged by the number of samples to obtain the impact value of the feature on the output, and the absolute value of the MIV of the 45 features is sorted to obtain the 10 features with the highest MIV value as a group of features, which completes the feature screening.

The XGBoost model, which is based on the Boosting framework, is a synthetic algorithm that combines basis functions and weights to fit the data, and is very powerful in missing value processing and prediction [4, 5]. In an effort to analyze the dataset in all aspects, we built the XGBoost model with a framework different from the previous random forest model, and the following is the specific process of model building [6].

First, for the 1200000 data in this problem, the XGBoost model can be represented as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \quad (i = 1, 2, \ldots 1200000) \tag{1}$$

where i denotes the predicted value of the i-th entry and k denotes the k-th decision tree.

The formula for the CART decision tree structure set is:

$$F = \{f(x) = w_{q(x_i)}\} (q: R^K \to \{1, 2, \ldots, T\}, w \in R^T) \tag{2}$$

ISSN: 2688-8645

where q is the tree structure of the sample mapping to the leaf nodes, T is the number of leaf nodes, and w is the real number fraction of leaf nodes. When constructing the XGBoost model, it is necessary to find the optimal parameters according to the principle of minimizing the objective function, so as to build the optimal model.

The objective function Obj can be divided into the error function L and the model complexity, with the following equation:

$$Obj = L + \Omega \tag{3}$$

$$L = \sum_{i=1}^{1200000} \left(y_i - \hat{y}_i\right)^2 \tag{4}$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5}$$

Where γ and λ are two hyperparameters to control the strength of the punishment.

For the XGBoost model, when optimizing the model using the training set, it is necessary to keep the original model unchanged and add a new function f to the model, as follows.

$$\begin{cases} \hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1\left(x_i\right) \\ \hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2\left(x_i\right) \\ \cdots\cdots \\ \hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t\left(x_i\right) \end{cases} \tag{6}$$

where $\hat{y}_i^{(t)}$ denotes the predicted value of the t-th model and $f_t(x_i)$ denotes the new function added to the t-th model. At this time the objective function can be expressed as

$$Obj^{(t)} = \sum_{i=1}^{1200000} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_t\left(x_i\right)\right)\right)^2 + \Omega \tag{7}$$

3. Results and Analysis

We derived the importance of its features by the MIV algorithm as shown in the Figure 1 below, where only the top ten important features are selected.

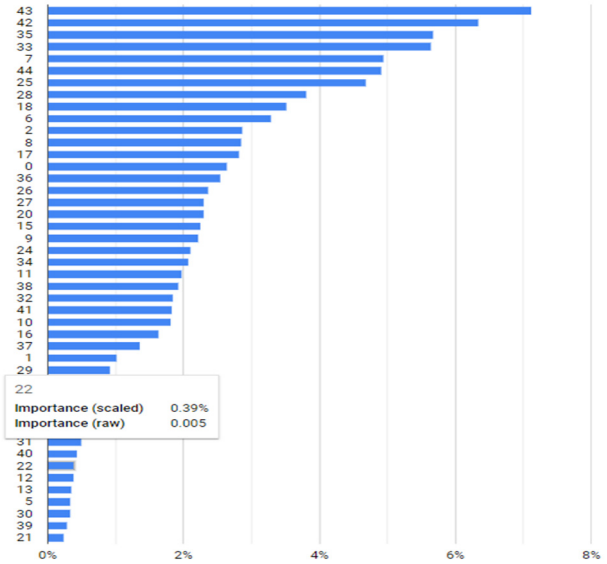

Figure 1. Feature Ranking

ISSN: 2688-8645

From the above graph, we pick the top 10 features: 44, 42, 35, 33, 7, 44, 25, 28, 18, 6.

After we selected 10 features, we substituted the preprocessed data of the corresponding features into the above XGBoost model for training, took 70% of the data as the training set, and 30% of the data as the test set, input 1,200,000 sets of data, and obtained the confusion of the XGBoost model. The confusion matrix of the Xgboost model is shown in Figure 2.

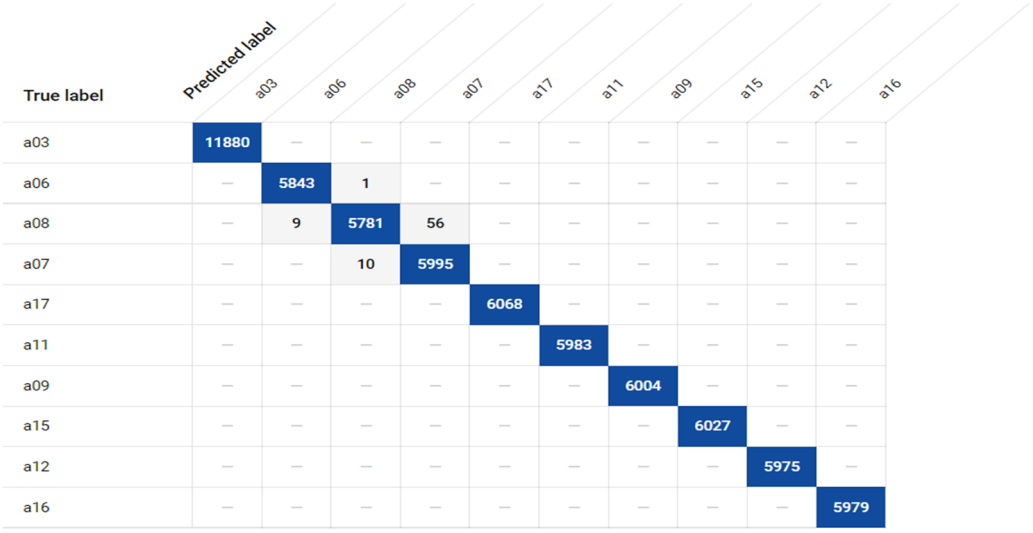

Figure 2. Confusion Matrix of the XGBoost model

The ROC curve of the XGBoost model, where the closer the curve is to the upper left corner, showing a right angle, the more effective the model is. From the Figure 3, it can be seen that the ROC curve of this model is close to the upper left corner, with an overall shape of a right triangle, and the value of AUC is 0.99.

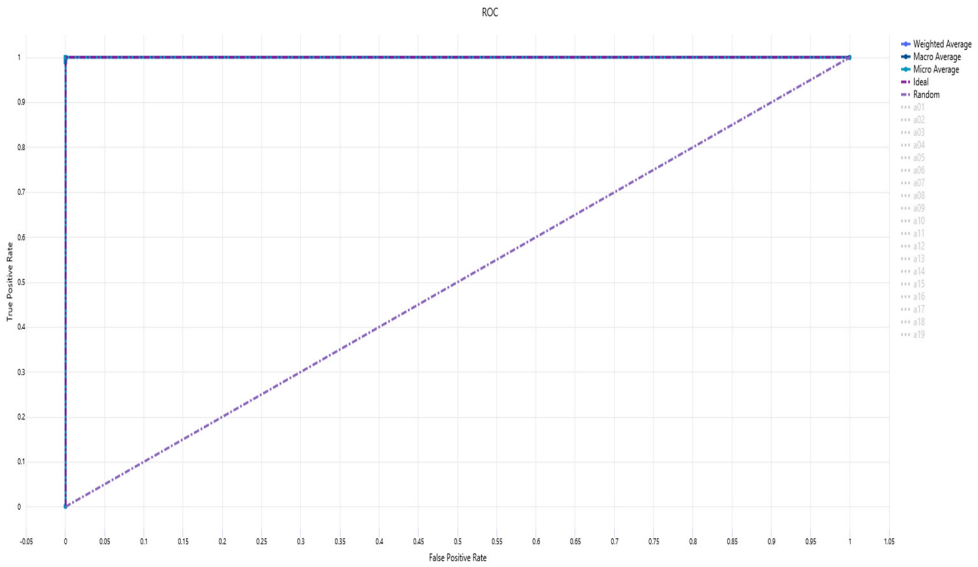

Figure 3. ROC curve of XGBoost model

3.1. Analysis of the Result

The precision (prediction accuracy), recall, f1-score and AUC values of the XGBoost model, are shown in the Table 1.

From Figure 1, it is easy to see that the top 10 features we derived by the MIV algorithm are 43, 42, 35, 33, 7, 44, 25, 28, 18, and 6 in order. After selecting the 10 features, the pre-processed data of the corresponding features are substituted into the above XGBoost model for training.

ISSN: 2688-8645

The prediction accuracy of the XGBoost classification model is close to 0.998, the recall is 1, and the F1 score is close to 1. In the test set of 1,200,000 data, only 76 data test errors can be seen from Figure 2, which means that the prediction ability of this model is nearly perfect and very suitable for solving human behavioral action classification problems.

Table 1. The evaluation value of each XGBoost model

Parameter Name	Value
Precision (prediction accuracy)	0.99
recall	1
f1-score	0.99
Roc_AUC	0.99

3.2. Improved Model

3.2.1. Oversampling Method

Due to the high cost of data, we need to make the model with good generalization ability with limited dataset. Therefore, we use SMOTE-Tomek integrated sampling method on the basis of the original data set to generate similar data samples on the basis of the original data and put them into the model trained with the original data for prediction evaluation.

The following is a brief description of the algorithmic principle of SMOTE-Tomek:

SMOTE-Tomek algorithm is an oversampling technique for synthesizing minority classes. The basic idea is to analyze the minority class samples and manually synthesize new samples to add to the dataset based on the minority class samples.

The main steps of the algorithm are as follows.

(1) For each sample x in the minority class, calculate the distance between that point and other sample points in the minority class to get the nearest k nearest neighbors (i.e., perform the KNN algorithm on the minority class points).

(2) Set the sampling ratio to determine the sampling multiplier, and for each minority class sample x, randomly select a number of samples from its k nearest neighbors, assuming that the selected nearest neighbor is x'.

(3) For each randomly selected nearest neighbor x', construct a new sample with the original sample respectively according to the following formula.

$$x_{now} = x + \text{rand}(0,1) * (x' - x) \tag{8}$$

3.2.2. Generalization Ability Test

Visualization method to observe the data distribution of each category of the original dataset (Figure 4).

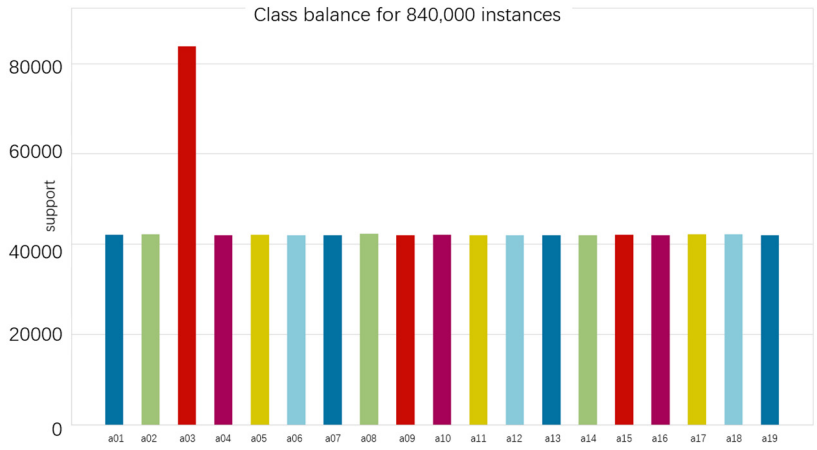

Figure 4. Original distribution

ISSN: 2688-8645

The SMOTE-Tomek integrated sampling method is used to generate similar data samples based on the training set of the original data, and the histogram of the data distribution after oversampling the original data (Figure 5).

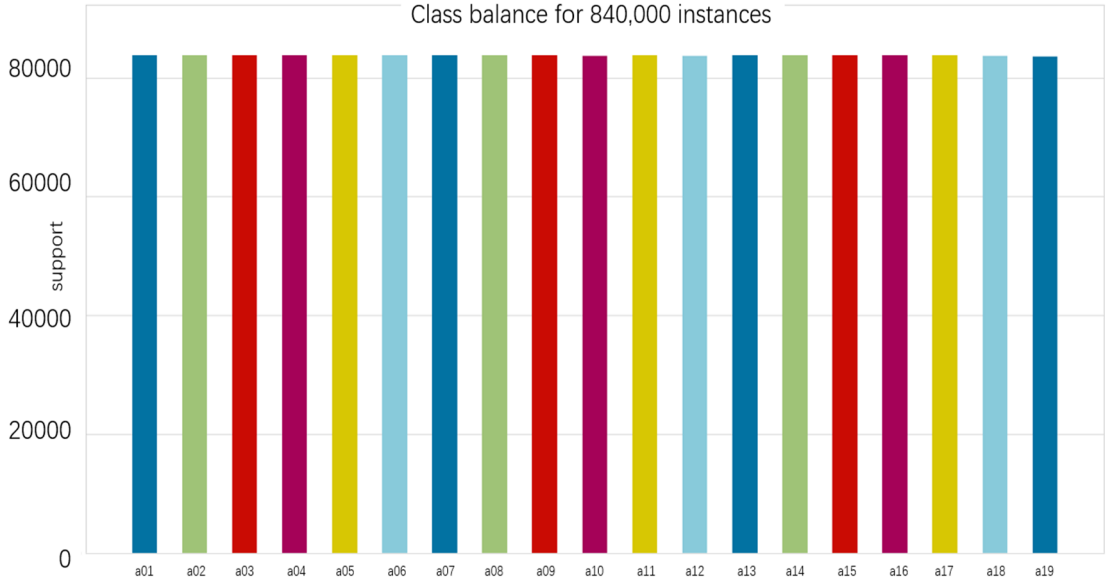

Figure 5. Post-processing data distribution

The precision, recall, f1-score and support of the XGBoost model obtained by inputting 1592183 sets of data are shown in the Table 2. From the table, we can see that the prediction accuracy of the model reaches 0.98, and the generated data also has an accuracy of 0.98 after substitution into the model, which can prove that the generalization ability of our model is good.

Table 2. The evaluation value of each XGBoost mode

Parameter Name	Value
precision	0.98
recall	0.99
F1 score	0.98
Roc_AUC	0.98

4. Conclusion

In this paper, by building the Xgboost classification model and using the MIV algorithm feature extraction for 1,200,000 data, and the simplified data set where this set of features is selected as the input to train the Xgboost classification model, the prediction accuracy of the Xgboost classification model is obtained as 0.99. For the problem of evaluating the generalization ability of the model, the SMOTE- Tomek integrated sampling method is used to generate similar data samples on the basis of the original data and put them into the model trained with the original data for prediction evaluation. The XGBoost classification model was then grid searched to obtain the optimal combination of parameters as learning_rate=0.5, max_depth=6, min_child_weight=5, n_estimators=100, and the parameters of the Xgboost classification model were set, and the original data set was used to train the integrated model using Xgb+Lgbm+RF as the base learner and logistic regression as the sublearners. The results show that the model is able to classify human behavior well with high accuracy rate by data from 10 sensors (i.e., 10 characteristics) and has good performance, which helps to classify people's behavior widely, and the model has good generalization ability.

ISSN: 2688-8645

References

[1] Wu W.H., Bui A.A.T., Batalin M.A., Liu D., Kaiser W.J. Incremental diagnosis methods for smart wearable sensor systems[J]. IEEE T. Information Technology, 2017, 11(5).

[2] Jafari Sadiqa, Byun YungCheol. XGBoost-Based Remaining Useful Life Estimation Model with Extended Kalman Particle Filter for Lithium-Ion Batteries[J]. Sensors, 2022, 22(23).

[3] S. KOCAOGLU, E. AKDOGAN. Comparison of Classification Algorithms for Detecting Patient Posture in Expandable Tumor Prostheses[J]. Advances in Electrical and Computer Engineering, 2020, 20(2).

[4] Tian Zhanxiao, Qu Wei, Zhao Yanli, Zhu Xiaolin, Wang Zhiren, Tan Yunlong, Jiang Ronghuan, Tan Shuping. Predicting depression and anxiety of Chinese population during COVID-19 in psychological evaluation data by XGBoost[J]. Journal of Affective Disorders, 2023, 323.

[5] Pan Zidong, Lu Wenxi, Wang Han, Bai Yukun. Groundwater contaminant source identification based on an ensemble learning search framework associated with an auto xgboost surrogate[J]. Environmental Modelling and Software, 2023, 159.

[6] MA H H. Application of random forest and XGBoost model in personal credit risk assessment[D]. The central university for nationalities, 2021. DOI: 10.27667 /, dc nki. Gzymu. 2021.000622.