ISSN: 1813-4890

Plant Seedling Classification Based on Ensemble Learning

Tingyu Li

College of Computer Science and Engineering, Northeastern University, Shenyang, 110819, China

20203465@stu.neu.edu.cn

Abstract

In this study, a method based on ensemble learning is proposed for the classification of plant seeds and seedlings. Classification of plant seeds and seedlings is an important task, which is of great significance to agricultural and biological research. However, due to the diversity and complexity of plant seedlings, traditional classification methods are often faced with the challenges of accuracy and robustness. In this study, we use ensemble learning method to improve the accuracy and stability of plant seed and seedling classification. Ensemble learning aims to produce more accurate final prediction results than a single classifier by combining the prediction results of multiple classifiers. We construct an ensemble model, which includes multiple classifiers, which are trained on different feature representations and classification algorithms. In order to realize ensemble learning, we adopt two key steps: First, we use feature selection algorithm to select the most informative feature from the original data. This is helpful to reduce the feature space and improve the effect of classifier. Secondly, we use multiple base classifiers, such as decision tree, support vector machine and random forest, and build the ensemble model by training different combinations of classifiers. Finally, the training data will be transferred to lightGBM as input data for the second training, and the final output will be the training result. We used a set of data sets containing images of multiple plant seeds and seedlings for experimental evaluation. Experimental results show that the method based on ensemble learning has better classification performance and robustness than single classifier. It can effectively deal with the diversity and complexity of plant seedlings and improve the accuracy and stability of classification. To sum up, this study proposes a method based on ensemble learning for the classification of plant seeds and seedlings. This method can effectively deal with the diversity and complexity of plant seedlings, and improve the accuracy and stability of classification. This has important application value in the fields of agriculture and biology research.

Keywords

Ensemble learning, classification, lightGBM, Plant Seedling Classification.

1. Introduction

1.1. Significance of research

The task of plant seedling classification based on machine learning has important practical significance and scientific value in the fields of agriculture and biology. The goal of this task is to accurately classify and identify plant seedlings by applying machine learning technology, so as to support agricultural professionals to better manage the crop growth process and improve the yield and quality of crops. In addition, accurate classification of plant seedlings is also helpful for biologists to study the growth and development mechanism of plants, understand the differences between different plant species and varieties, and promote scientific research in ecology and botany.

ISSN: 1813-4890

At present, the research in this field mainly focuses on the following aspects: First, establish a large-scale plant seedling image data set to provide rich and diverse sample data for plant seedling classification tasks. Secondly, select and extract the most discriminant and informative features, so as to effectively represent the morphological and structural features of plant seedlings and reduce the influence of redundant and noisy information. For feature representation, researchers use various feature extraction algorithms and feature selection methods, such as local binary pattern (LBP), color histogram, texture features and so on. Finally, a variety of machine learning algorithms are applied to plant seedling classification tasks, including traditional decision tree, support vector machine (SVM), random forest and other methods. In this study, the concept of ensemble learning is mentioned, which connects single machine learning models to improve the accuracy and robustness of plant seedling classification.

1.2. Solutions

Firstly, the data set is processed to extract the feature matrix of each image. Then it is trained in each machine learning model and debugged by gridsearchcv. The parameters maxdepth and num leaves which are the most important to improve the accuracy are preferentially selected, and then the parameters min child samples, min child weight and learning rate are debugged to prevent over-fitting. Finally, the training results and prediction results between single models, single models and integrated models, and different integrated models are compared, and the final integrated learning model is selected according to the final accuracy.

1.3. Value

The task of plant seedling classification based on machine learning has important scientific and practical value. Applying machine learning algorithm to accurately classify plant seedlings can provide effective agricultural management strategies for agricultural production, improve crop yield and quality, and promote the sustainable development of agriculture. In addition, the classification of plant seedlings is also helpful to ecological research, revealing the ecological preferences and habitat needs of different plant species, and promoting the conservation of biodiversity and the maintenance of ecological balance. In biological research, plant seedling classification provides important data and information for plant growth and development, and promotes scientific progress in plant genetics, evolutionary biology and population biology. In addition, the task of plant seedling classification also plays an auxiliary role in crop breeding and improvement, and provides an important reference for breeding better crop varieties. Therefore, the research and application of plant seedling classification task based on machine learning has wide value, and brings important scientific and practical achievements to agriculture, ecology and biology.

2. Related work

2.1. Data processing

Firstly, the original data set is equalized by histogram. Then the image is transformed into HSV format to extract the leaf part, and the SIFT, HOG and LBP feature matrices of each image are mosaic.

2.1.1 Histogram equalization

The b, g, r of the image are extracted by using cv2. split () function, then the third one is equalized by using cv2. equalizeHistory () function, and finally the final image is combined by using cv2. merge ().

2.1.2 Extraction of leaf parts

ISSN: 1813-4890

Delineate the upper and lower bounds of colors to be extracted. Here we choose green and cyan. Firstly, the equalized image is filtered by Gaussian filtering, and then the image is converted into HSV format, because HSV can divide each color into a range, but RGB format, each color can not be divided into a specified range. Then, the mask image is obtained, and the pixel value within the threshold value is set to white, and the pixel value outside the threshold value is set to black, and then the mask image and the processed image are taken bit by bit to obtain a leaf part.

2.1.3 SIFT characteristics

Firstly, we use SIFT.detectAndCompute (image, None) function to obtain SIFT key point features, and then grayscale the image to be processed. Then we use kp to store key point information, and des expands kp information into a matrix composed of 128 features.

SIFT algorithm steps: 1. Scale space extremum detection 2. Key point location 3. Key point direction parameters 4. Key point description 5. Key point matching.

2.1.4 BOW+K-means

Firstly, all the images are divided into many small patches, which are broken up by k-means and divided into a similar group, that is, the combination of visual words with similar meanings. As the basic words in the dictionary, the SIFT algorithm is used to extract visual word vectors from different types of images. These vectors represent local invariant feature points in the images, and these points can be replaced by synonyms in the dictionary. Finally, the whole image data set is scanned to count the number of times each word in the word list appears in the image, thus representing the image as a K-dimensional numerical vector.

In our project, firstly we use cv2. BOWKMeansTrainer () to create a BOW trainer, and specify the number of words in the final vocabulary dictionary to be 100. Then we add SIFT key point features extracted from all pictures to the BOW trainer, then we carry out K-means clustering to get the cluster center, and define FLANN matching algorithm. Here we choose to use kd tree nearest neighbor search. Finally, we use the vocabulary dictionary and FLANN matching algorithm to initialize BOW extractor and use it as the return value of the function.

2.1.5 HOG characteristics

First, the images to be extracted with HOG features are changed to the size of (128, 128) in order to ensure that the number of HOG features extracted from each image is the same. We need to pay attention to the shape and texture on the image. In order to observe the spatial distribution of these gradients, we need to divide the image into grids and calculate multiple histograms. Firstly, the image is divided into small connected regions, and then the gradient or edge direction histogram of each pixel in the cell unit is collected. Finally, these histograms can be combined to form a feature descriptor. The core idea is that the detected local shape can be described by the distribution of gradient or edge direction. HOG can capture the local shape information well and has good invariance to geometric and optical changes.

2.1.6 LBP characteristics

First, the images to be extracted with LBP features are changed to the size of (128, 128) in order to ensure that the number of LBP features extracted from each image is the same. LBP is an operator used to describe the local texture features of an image. It has the advantages of rotation invariance and gray invariance. The principle of LBP is to take the center pixel of the window as the threshold, and compare the gray values of eight adjacent pixels with them. If the surrounding pixel value is greater than the center pixel value, the position of the pixel point is marked as 1, otherwise it is 0. In this project, we use circular LBP operator, which extends the 8 * 8 neighborhood to any neighborhood, and replaces the square neighborhood with circular neighborhood. The improved LBP operator allows any number of pixels in the circular neighborhood with radius R. Thus, LBP operators with P sampling points in a circular region with radius R are obtained.

ISSN: 1813-4890

2.1.7 Fill in missing values

Only in LBP feature, we first reduce the dimension of the LBP feature matrix, then transform it into DataFrame format, and then use fillna function to change the missing value into 0.

2.1.8 Standardization

The standard score of normalized feature sample x by removing the mean and scaling it to unit variance is calculated as z = (x-u)/s, where u is the average of the training sample, 0 if with_mean = False, s is the standard deviation of the training sample, and 1 if with_std = False. By calculating the relevant statistical information of the samples in the training set, each feature is centered and scaled independently. The mean and standard deviation are then stored to be used on later data using transformations.

2.1.9 Dimension reduction

We use PCA principal component analysis to reduce the dimension, which is a means of feature selection (reconstruction). The original feature space is mapped, so that the new mapped feature space data are orthogonal to each other, and the distinguishable low-dimensional data features are preserved as much as possible. In the code, estimator.explained_variance_ratio_ indicates that each dimension can represent the proportion of the original features after dimensionality reduction, traversing and accumulating them for output, which is convenient to determine a good dimensionality reduction degree, so that there are not many dimensions and it can represent a higher proportion of the original features, and minimize the feature loss caused by dimensionality reduction.

2.2. Model Training

Using gridsearchcv to debug some parameters of LightGBM, the most important parameters for improving accuracy, maxdepth and num leaves, are selected first, and then the parameters, min child samples, min child weight and learning rate, are debugged to prevent over-fitting.

2.2.1 LightGBM

LightGBM model is improved on the basis of GBDT to solve the problem of large consumption of time and space resources. The decision tree algorithm based on Histogram is used to discretize the continuous floating-point features into k integers, which makes the storage more convenient and the computational cost less. At the same time, it can save a lot of space and time overhead by reducing samples by unilateral gradient sampling.

2.2.2 RandomForest

Random forest is a classifier containing multiple decision trees, and its output is determined by the mode of categories. Random forest includes two randomness, sample randomness and feature randomness. Firstly, m training sets are generated by bootstrap method, and a decision tree is generated for each training set. When splitting, some features are randomly selected from the features, and the final prediction results are the most selected from the m results.

2.2.3 GBDT

GBDT adopts the strategy of multi-model integration, and fits the residual error, so as to reduce the deviation and variance of the model. GBDT is an additive model, which sums up the predicted values of all basic models as the final predicted value. The basic model is a series structure, and its prediction rules can be obtained by reasoning and updating the model every time:

$$\widehat{y_n} = \sum_{k=1}^{K} T_k(x_n)$$

ISSN: 1813-4890

2.2.4 SVC

SVM is an algorithm to solve the binary classification problem. SVM is used to divide samples into two categories. The principle of segmentation is to find an optimal decision boundary to maximize the segmentation interval, and the support vector is the nearest point to the optimal boundary. According to Lagrange multiplier method, KKT condition and dual problem, a set of optimal w values of SVM objective function are calculated according to the following steps.

$$min\frac{1}{2}\|w\|^2\ s.t\ y\cdot(w^T\cdot x_i+b)\geq 1\ ,i=1,2,\dots,n$$

2.2.5 SGD

SGD is a stochastic gradient descent method in deep learning, and it is a strategic problem of parameter optimization. The goal of the algorithm is to find the model parameters that minimize the error of the model on the training set. By modifying and perfecting the model parameters, it can be carried out along the direction of error gradient reduction until the minimum error value is found. It is a first-order optimization algorithm, which only considers the first derivative of the function, and randomly selects a data to calculate when calculating the fastest descent direction, so as to speed up the iteration.

2.2.6 ExtraTrees

Using the top-down divide-and-conquer strategy, the core is how to choose the best attribute in the recursive process. At the beginning, the root node is constructed, all the training data are put into the root node, the optimal attribute is selected, the sub-data sets are divided, and recursive processing is carried out in turn. If the stop condition is met at last, the leaf node is generated, and when all the subsets are assigned to the leaf node, a decision tree is generated.

There are three stopping conditions:

A. All the samples in the current node belong to the same category, so there is no need to divide them;

b. The current attribute set is empty, or all samples have the same value on all attributes, so it is impossible to divide;

c. The sample set of the current node is empty and cannot be divided.

In addition, we need to use pruning method to prevent over-fitting of decision tree, and actively remove some branches to reduce the risk of over-fitting. The basic strategies are pre-pruning and post-pruning. Pre-pruning terminates the growth of some branches in advance by judging the precision of verification set before and after partition. Post-pruning is to judge whether the accuracy of verification set is reduced from the deepest node, and then "turn back" pruning. Pre-pruning can reduce the training time and test time, and reduce the risk of over-fitting, but it will lead to easy under-fitting; After pruning, the training time is increased and the test time is shorter, which reduces the risk of over-fitting and does not increase the risk of under-fitting. Moreover, the generalization of the model is higher.

2.3. 2.3 Parameter tuning

For some prediction models, we have optimized the parameters, using the principle of coarse adjustment first and then fine adjustment. Using gridsearchcv to debug some parameters of lightgbm, the most important parameters to improve accuracy are max_depth and num_leaves. The first debugging time is 7.58 s and the prediction accuracy is 87.5% when max_depth=2 and num_leaves=12; After debugging, we choose max depth=4 and num leaves=2. At this time, the prediction accuracy and time are improved.

ISSN: 1813-4890

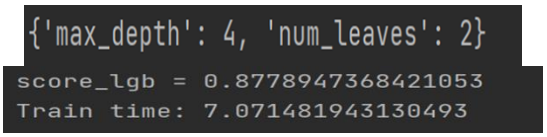

Fig.1. max_depth versus num_leaves tuning

Then debug the parameters min_child_samples and min_child_weight to prevent over-fitting, and get higher accuracy when min_child_samples = 18 and min_child_weight=0. 001.

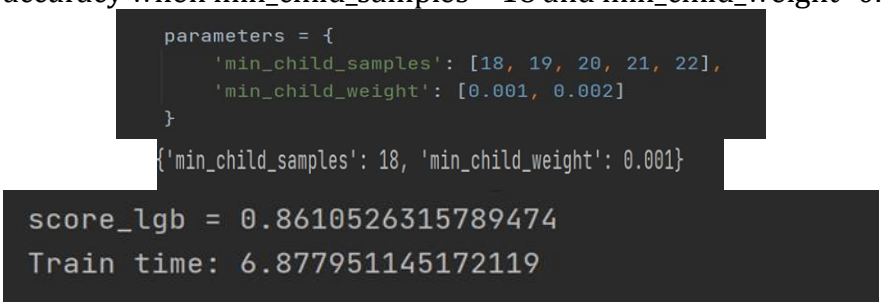

Fig.2. Min_child_samples versus min_child_weight=0. 001 tuning1

The number of iterations n_estimators first sets the range to 200-800, and then gradually narrows the range, which is optimal when n_estimators = 500.

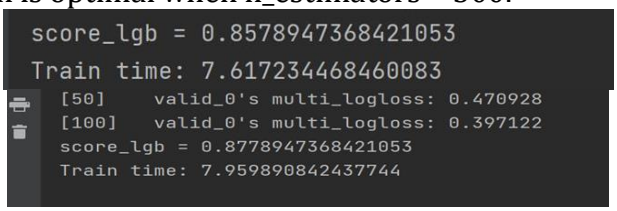

Fig.3. n_estimators tuning2

The model works well where the learning rate learning_rate is 0.09.

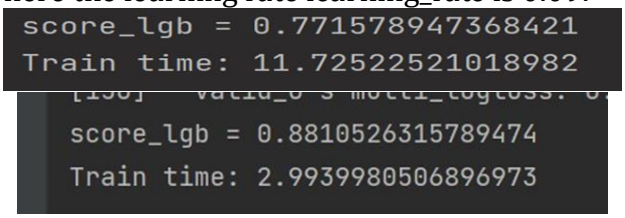

Fig.4. learning_rate tuning

3. Overall process

First, I equalize the images in the original data set by histogram. In the aspect of data processing, firstly, the color range of green and cyan (HSV image in this case) is delineated, and the leaf part in each image is extracted by opencv. Next, SIFT features are extracted from each image by SIFT + BOW + K-means, and then HOG and LBP features are extracted from each image. It should be noted that the images need to be changed into (128, 128) sizes in advance when extracting the latter two features, so as to ensure that the number of HOG and LBP features extracted from each image is the same. Then the SIFT feature matrix is normalized, the HOG feature matrix is normalized and dimensionality reduced, the LBP feature matrix is normalized and dimensionality reduced, the missing values are filled, normalized and dimensionality reduced again, and the three feature matrices after processing are spliced. Finally, using the StratifiedShuffleSplit () function, the data set is divided into verification set and training set by class.

After data processing, training is carried out. In this experiment, I use several prediction models, deep learning models and integrated model training data respectively, and get preliminary training results through pre-training. Some parameters of lightgbm are debugged by gridsearchcv, and the most important parameters for improving accuracy are maxdepth and

ISSN: 1813-4890

num-leaves, and then the parameters min child samples and min child weight to prevent over-fitting are debugged. The iteration times n-estimators and learning rate are also adopted in the same way. Then compare the training results and prediction results between single models, single models and integrated models, and different integrated models. Finally, choose lightGBM, RandomForest, SVC, SGD and ExtraTrees as independent basic functions, and transfer the training data to lightGBM as input data for the second training, and take the final output as training results.

4. Comparative experiment

4.1. Histogram equalization

Advantages:

Using histogram equalization can enhance the contrast of images with small dynamic range, transform the histogram of the original image into a uniform distribution form, thus increasing the dynamic range of the difference between pixels, making the image contrast and clarity larger, and achieving the effect of enhancing the image.

4.2. Extracting Leaf Part from Image

Advantages:

a.The research is the classification of plant seedlings, so the leaves of plants should be used. However, because the original image contains useless information such as soil and Stone except plant leaves, the leaves of plants should be extracted before feature extraction, which is more conducive to subsequent model training.

b.We use Gaussian filtering to process the image before extracting the leaf part. Gaussian filtering can eliminate the Gaussian noise produced or mixed in the image during digitization and make it become linear and smooth.

4.3. SIFT Features

Advantages:

a.SIFT feature is a local feature of image, which is invariant to rotation, scale scaling and brightness change, and stable to viewing angle change, affine transformation and noise to a certain extent;

b.Good discrimination, abundant information, suitable for fast and accurate matching in massive feature database;

c.It has multi-quantity, even a few objects can produce a large number of SIFT feature vectors;

d.It has high speed, and the optimized SIFT matching algorithm can even meet the real-time requirements;

e.It is extensible and can be easily combined with other feature vectors.

Disadvantages:

a.SIFT is highly dependent on the gradient of pixels in a local area, and it is possible that this area is not properly obtained, which leads to the inaccurate main direction we find, which leads to the large error of the calculated feature vector, thus failing to match successfully;

b.If the distribution of pixel values is too concentrated, SIFT will not perform very well.

4.4. BOW + K-means

Advantages: The algorithm is relatively simple, easy to understand and effective.

Disadvantages:

a.When the image is represented by visual words, the position information of image features is not included;

ISSN: 1813-4890

b.The dictionary is too large, which easily leads to the problem of sparse data.

4.5. HOG characteristics

Advantages:

a.Keeping good invariance to both geometric and optical deformation of the image;

b.For the feature extraction of rigid objects, it has good characteristics.

Disadvantages:

a.The characteristic dimension is large;

b. Large amount of calculation;

c. The algorithm cannot deal with occlusion problem.

4.6. LBP Features

Advantages:

a.Rotation invariance and gray invariance;

b.The operation speed of the algorithm is fast.

Disadvantages: Sensitive to direction information.

4.7. LightGBM

The effect is the best: compared with other models, the speed is faster and the accuracy is higher.

Advantages:

a.The speed is improved: the decision tree algorithm based on Hostogram is adopted to discretize the continuous floating-point features into k integers, which makes the storage more convenient, the calculation cost less and the time complexity greatly reduced; In the training process, unilateral gradient sampling is used to reduce samples, which can save a lot of space and time overhead; The growth strategy based on Leaf-wise algorithm is used to construct the tree, which can prevent over-fitting and reduce a lot of unnecessary computation;

b.The space occupancy rate is reduced: XGBoost needs to use extra space to record the index of eigenvalues and their corresponding sample statistics after using pre-sorting, while LightGBM uses histogram algorithm to transform eigenvalues into bin values, and does not need to record the index of features to samples, which greatly reduces memory consumption; In the training process, the algorithm of mutual exclusive feature bundling is used to reduce the number of features, which not only achieves the purpose of reducing the number of features, but also reduces the memory consumption.

Disadvantages:

a.Deep decision trees may grow, resulting in over-fitting. Therefore, LightGBM adds a maximum depth limit on Leaf-wise to prevent over-fitting while ensuring high efficiency;

b.LightGBM is an iterative algorithm, and each iteration is adjusted according to the previous prediction results, so LightGBM is sensitive to noise;

c.When searching for the optimal solution, it is based on the optimal segmentation variable, and does not consider the idea that the optimal solution is the synthesis of all characteristics.

4.8. RandomForest

The effect is inferior to LightGBM: the accuracy is slightly lower than LightGBM, but the speed is

Soon.

Advantages:

a.Use two randomness (sample randomness and feature randomness);

b.Being able to process large amounts of data and still maintain high accuracy;

ISSN: 1813-4890

c.In determining categories, the importance of variables is assessed.

Disadvantages:

a.Random forests will be over-fitted in some noisy classification or regression problems;

b.For the data of attributes with different values, attributes with more values will have greater influence on random forest.

4.9. GBDT

Advantages:

a.More flexibility: Flexibility to handle various types of data, including continuous and discrete values;

b.High accuracy: In the case of relatively less parameter adjustment time, the prediction accuracy is also relatively high, compared with SVM;

c.Strong robustness: Using some robust loss functions, the robustness to outliers is very strong.

Disadvantages: Because of the strong dependence between weak learners, it is difficult to train in parallel. Partial parallelism can be achieved by self-sampling SGBT.

4.10. SVC

Advantages:

A. It is a good prediction generator, which provides over-fitting, noise data and outlier processing;

b. The training speed is faster;

c. It can automatically detect data nonlinearity without variable transformation;

Disadvantages: It is better to deal with the binary classification problem, but it is not very good to other problems.

4.11. SGD

Advantages: SGD is iteratively updated once through each sample, and the training speed is fast.

Disadvantages: SGD has more noise than BGD, so it is not carried out in the direction of overall optimization every iteration, and its accuracy will decrease, which is not global optimization and is not easy to be implemented in parallel.

4.12. ExtraTrees

Advantages:

a.High speed, relatively small amount of calculation and easy conversion into classification rules;

b.High accuracy, high accuracy of the classified rules, can clearly display important fields;

c.There are no parameters, so no parameter setting is needed.

Disadvantages:

a.Because of depth-first search, it will be limited by memory size and cannot handle big data;

b.The improvement for processing large data sets increases the overhead of classification and reduces the accuracy of classification. When the number of categories increases, the error rate will rise, and the data with time sequence restriction needs to be preprocessed.

5. Experimental results

Training results:

ISSN: 1813-4890

```
Train Epoch: 2 [320/3325 (10%)] Loss: 0.486977
Train Epoch: 2 [640/3325 (19%)] Loss: 0.397502
Train Epoch: 2 [960/3325 (29%)] Loss: 0.442869
Train Epoch: 2 [1280/3325 (38%)]    Loss: 0.150242
Train Epoch: 2 [1600/3325 (48%)]    Loss: 0.110013
Train Epoch: 2 [1920/3325 (58%)]    Loss: 0.302237
Train Epoch: 2 [2240/3325 (67%)]    Loss: 0.181905
Train Epoch: 2 [2560/3325 (77%)]    Loss: 0.381454
Train Epoch: 2 [2880/3325 (87%)]    Loss: 0.090340
Train Epoch: 2 [3200/3325 (96%)]    Loss: 0.279651
epoch:2,loss:0.2917152776096303
1425 45

Val set: Average loss: 0.3036, Accuracy: 1281/1425 (90%)
```

Fig.5.VGG model training results

```
[50]    valid_0's multi_logloss: 0.661156
[100]   valid_0's multi_logloss: 0.51623
[150]   valid_0's multi_logloss: 0.457002
[200]   valid_0's multi_logloss: 0.425661
[250]   valid_0's multi_logloss: 0.408761
score_lgb = 0.8757894736842106
Train time: 7.940427303314209

进程已结束,退出代码0
```

Fig.6. LightGBM training results

```
[50]    valid_0's multi_logloss: 0.661156
[100]   valid_0's multi_logloss: 0.51623
[150]   valid_0's multi_logloss: 0.457002
[200]   valid_0's multi_logloss: 0.425661
[250]   valid_0's multi_logloss: 0.408761
score_lgb = 0.8757894736842106
Train time: 7.940427303314209

进程已结束,退出代码0
```

Fig.7.RandomForest training results

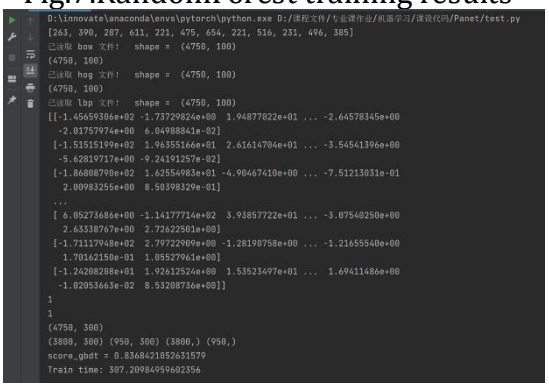

Fig.8.GBDT training results

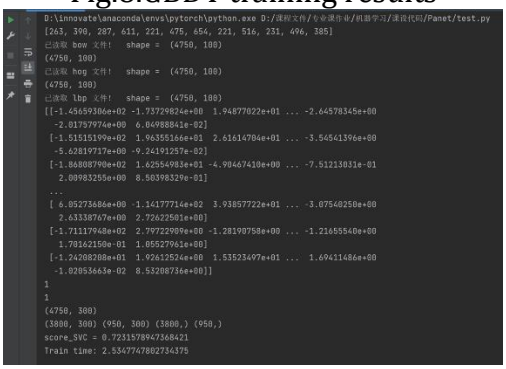

Fig.9.SVC training results

ISSN: 1813-4890

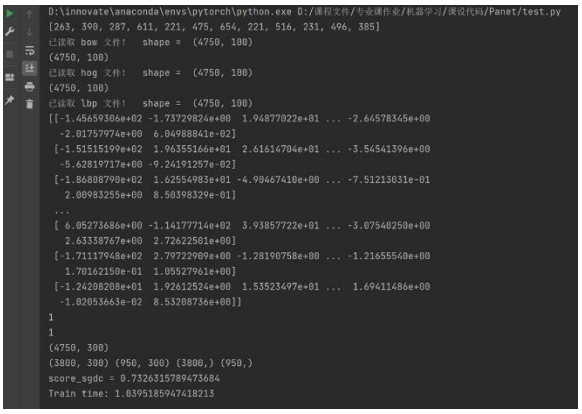

Fig.10.SGD training results

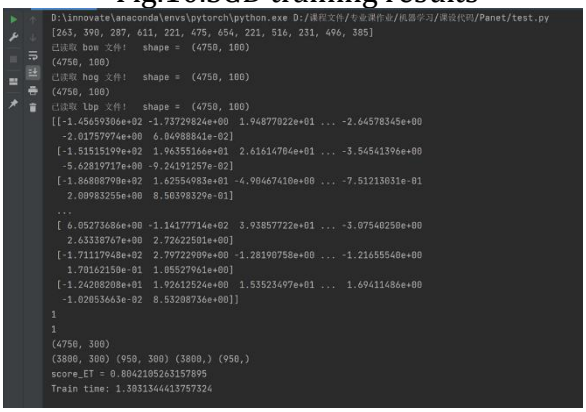

Fig.11. ExtraTrees training results

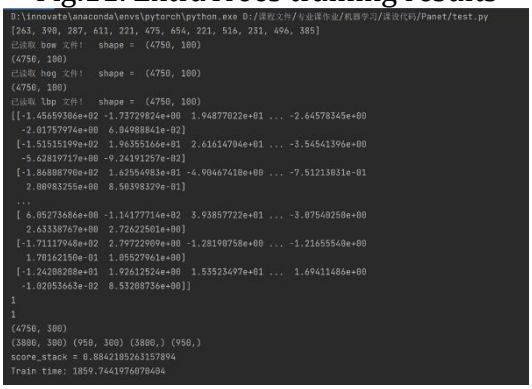

Fig.12. LighGBM, random forest, GBDT integration (LighGBM as meta-function)

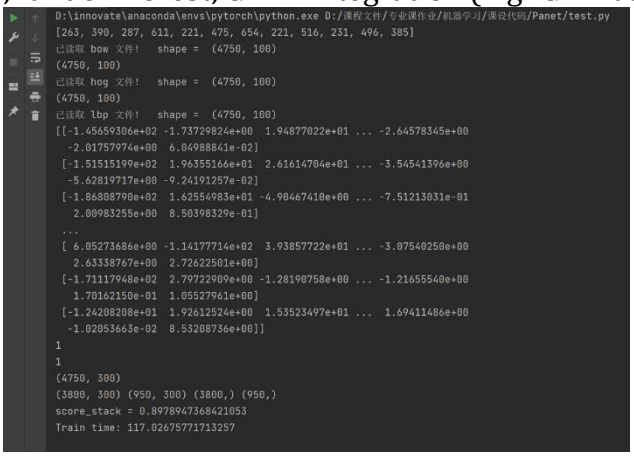

Fig.13.LighGBM, Random Forest, SVC, SGD, ExtraTrees (SVC as Meta Function)

ISSN: 1813-4890

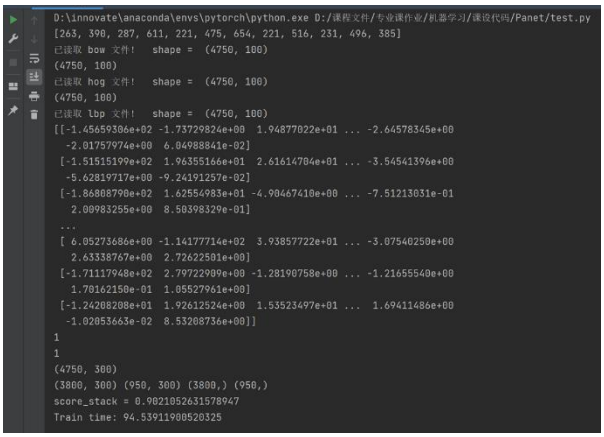

Fig.14. LighGBM, Random Forest, SVC, SGD, ExtraTrees (LighGBM as Meta Function)

From the experimental results, we can see that the accuracy of lightgbm is 87.57% and the time is 7.94 s; The accuracy rate of random forest is 82.84% and the time is 2.06 s; GBDT has a relatively long time, but its accuracy is similar to that of random forest; The accuracy of SVC is relatively low, 72.31%, and the time is 2.53 s; The accuracy of SGD was 73.3% and the time was 1.04 s; Finally, the limit tree, whose accuracy is 80.4% and the time is 1.30 s. Then we try to use the integration method to improve the accuracy of the model. The first time we use LightGBM, random forest, GBDT as independent basic functions, and output the training data to LightGBM for the second training. The result is 88.4%. The time is more than 30 minutes. Then we remove the time-consuming GBDT and add SVC, SGD, ExtraTrees. We use SVC as meta-function, and continue training to get 89.78%. The visible accuracy is improved in 117s. Finally, through continuous attempts, we choose LightGBM, RandomForest, SVC, SGD, ExtraTrees as basic functions and LightGBM as meta-function to get 90.21% accuracy The time is 94.5 s. Through the above comparison, LightGBM performs best in a single model, and the following four points are summarized through analysis:

a.Based on the decision tree algorithm of ∼, the continuous floating-point features are discretized, which makes the memory occupy space, the calculation cost is less and the speed is faster;

b.Reducing samples and saving time by unilateral gradient sampling;

c.Mutually exclusive feature binding to reduce the number of features and achieve the purpose of dimensionality reduction;

d.Leaf growth strategy with depth limitation can effectively prevent over-fitting and ensure the accuracy of test.

The second place is random forest, which is second only to LightGBM in accuracy, but it is very fast, and summarizes the following three points:

a.Using two randomness can avoid over-fitting;

b.Able to process a large amount of data and still maintain high accuracy;

c.When deciding the category, the importance of variables will be evaluated to ensure the accuracy of the results.

References

[1] Fu X T, Chang Q R, Zhang Y M et al.Estimation of chlorophyll content in kiwifruit leaves based on Stacking ensemble learning [J]. Agricultural Research in Arid Areas, 2023, 41 (04): 247-256.

[2] Rong Jiguang, Liu Weigang, Tian Jubo et al.A tactical target classification method based on ensemble learning [J]. Radio Engineering, 2023, 53 (07): 1700-1705.

[3] Yang Zekun, Zhu Mengyuan, Zhou Mingqian.Research on battery capacity prediction method based on LightGBM [J]. Post and Telecommunications Design Technology, 2023 (07): 70-74.

[4] Huang Jinming, Du Mont.Building a used car price prediction model based on Stacking ensemble learning [J]. China Science and Technology Information, 2023 (14): 88-89.

[5] Li Haixin, Zhang Jiaojiao, Wang Yu et al.5G user prediction and 5G network planning based on LightGBM algorithm [J]. Information and Communication Technology, 2023, 17 (03): 69-74.