Candidate Number :
253251
3 June 2022

Machine Learning Report

Sunny or Not Sunny—A Binary Classifier Problem

The assessment reveals that this is a Binary classification in typical supervised learning.

1 Approach

Even though there are many models provided by the machine learning library, I have chosen logical regression with high explanatory degree and random forest and support vector machines that can rival the accuracy of neural network models.

1.1 Logistic Regression

Logistic regression is also called logarithmic probability regression. Although the algorithm name is called logistic regression, it is a classification algorithm because logistic regression uses a method similar to regression to solve the classification problem.

1.2 Random Forest

Random Forest is an ensemble learning algorithm for classification and regression and it generates a multitude of decision trees classifies based on the aggregated decision of those trees. Then the objects are classified in turn. Finally, the classification information of each decision tree is summarised. The mode category in all prediction categories is the object category predicted by random forest, which greatly improves the prediction accuracy.

1.3 Support Vector Machine

SVM, also known as support vector machines, is a non-probabilistic binary linear classifier. SVM is a class of classifiers that perform binary classification based on supervised learning. Its purpose is to find a hyperplane to segment the samples. The segmentation principle is to maximise the interval, and finally transform it into a convex quadratic programming problem to solve.

2 Methodology

2.1 Data Description

The data set given for training have 256 samples which are lower numbers for machine learning training and have complete values. In contrast, the additional training data set have 2331 samples with missing some feature.

Table 1

	Sunny	Not sunny	Samples	Positive Rate	Missing Rate
training	146	113	259	56.3%	0
additional_training	1409	922	2331	60.4%	20%
training+additional	1555	1035	2590	60%	18%
training+additional(confidence=1)	450	570	1020	44%	18%

Training Set Compared

Using fewer samples training set may cause overfitting of the trained model. Additional training set may have been an important factor in improve the model performance. As discussed above, this report will base on the training and additional training set with confidence equals 1 which has a proper samples and enough information for training model. And this data set already have feature extraction by CNN features concatenates GIST features.

2.2 Data Pre-processing

1.Dealing with missing values. Handling with missing values may be divided into several groups[1].:

a) Delete it—will cause our training set back to original
b) Impute missing value with Mean/Median/Most_frequent —this is common use in imputing the missing value
c) Predict the missing value—will consume a lot of computation. Sklearn provides two common method IterativeImputer and KNNImputer.

Using KNNImputer in my project seems provide a more robust and efficient way to save more accurate information, thereby I use it for imputing missing value.

2.Feature fusion and Feature reduction

The data set has completed the simplest concatenate feature fusion. It is not yet clear whether concatenate feature fusion is made worse by individual CNN features or GIST features. For this reason, I did many experiments. The results show that even with the removal of the features of the GIST, our model is still accurate and generalised.

So at the beginning, I deleted the feature dimension of GIST and reduced the dimension of CNN by principal component analysis.

2.3 Data Splitting

In summarise our data set characteristics :

a) Almost the balance positive and negative ratio

b) Not a large sample size

Therefore, my data is divided by the training set and the test set with 80 percent and 20 percent, and 5fold-Cross validation is applied to the training set for a balance of computational performance and sample usage.

2.4 Trained Model

a) Model Choose

As discussed above, the training of the model mainly uses the standard model (Logistic Regression) /(Random Forest)/ (SVM) + 5 fold Cross-Validation

Model Optimisation

b) GridSearchCV(for find the best parameters)

It is python's parameter auto-search module. We just need to tell it what parameters it wants to tune and what range it can take, and it will just run away and tell us which parameter is the best.

2.5 Evaluate Model Performance

To assess the model performance, the evaluation of generalisation in new data was need. For this reason, I use the performance on the validation set as an approximation of the generalisation performance. There are many metrics below to evaluate performance. Back to task in this assessment, we need to improve the accuracy of sunny forecasting as much as possible.

a) Accuracy

Accuracy refers to the proportion of model predicting the correct results and it is the most commonly used indicator in classification problems. However, Accuracy is not enough indicator for data sets.

b) Precision and Recall

Precision means that the percentage of samples divided into positive examples is actually positive. Recall means that the percentage of practical samples are divided into actually cases.

c) F1-Score

$$\frac{1}{F_1} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right)$$

Generally speaking, precision and recall are a pair of contradictory measurements sometimes. F1-score is a combination of Precision and Recall. Thus, in the project, f1 is used to evaluate the main indicators of performance[2].

d) Receiver Operating Characteristics and AUC

The horizontal coordinate of the ROC curve was false positive rate (FPR), Vertical Positive Rate (TPR). It receives less impact from the dataset than the precision-recall curve. This report will therefore evaluate the model by drawing ROC curves.

3 Result and Discussion

Table 3-1

	Accuracy	Recall	Precision	F1_Score	AUC	best_validation score	worst_validation_score
Logistic Regression	0.833	0.569	0.852	0.832	0.91	0.802	0.565
Random Forest	0.799	0.64	0.711	0.803	0.85	0.756	0.734
SVM	0.843	0.589	0.851	0.848	0.91	0.803	0.590

Trained By CNN features

Table 3-2

	Accuracy	Recall	Precision	F1_Score	AUC
Logistic Regression	0.794	0.668	0.772	0.758	0.87
Random Forest	0.593	0.528	0.582	0.497	0.61
SVM	0.804	0.678	0.776	0.777	0.88

Trained By GIST features

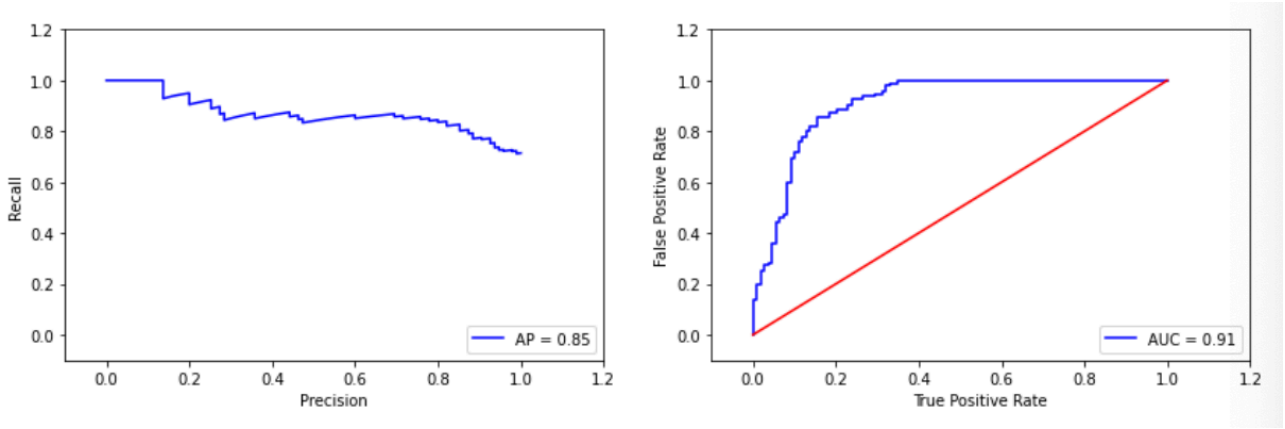
Table 3-3

	Accuracy	Recall	Precision	F1_Score	AUC
Logistic Regression	0.843	0.581	0.880	0.837	0.93
Random Forest	0.819	0.610	0.768	0.822	0.87
SVM	0.843	0.600	0.852	0.845	0.91

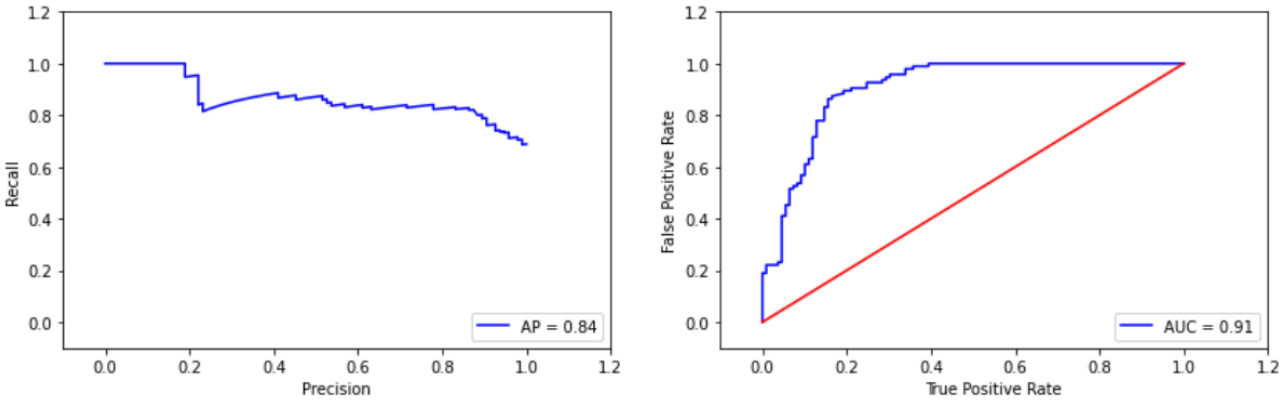
Trained By CNN+GIST features

From the results, there is little difference between the results of training only using CNN features and the results of training with complete features. The effects of LogisticRegression on classification are similar to those of Support Vector Machine. Although both SVM and logistic regression models can get high scores on my training set, they are too easy to fit.

Also, I plotted the P-R curves and ROC of Logistic Regression Model and Support Vector Machine.



Pic3-1. P-R curve, and ROC (Trained by Logistics Regression with CNN features)



Pic3-2. P-R curve, and ROC (Trained by SVM with CNN features)

Apply our best trained RandomForest model which has the best robust to testing. From the result data in "results.csv", we can see that the positive number is 1503 and the negative number is 1315. The positive proportion was 53.406%, and the negative proportion was 46.6%. In the SVM model I trained, the positive and negative ratios obtained are completely opposite to the given positive and negative ratios. I guess it is over fitting in the training process.

In this project, my main focus is on the selection of interpolation methods and the fusion of features. First, in terms of interpolation methods, KNN, although not the best in terms of performance and accuracy, achieves a balance in performance. Secondly, k-cross fold validation and GridSearchCV () are used to make the best choice in parameter adjustment of the model. Finally, and most importantly, feature fusion above I hope to better fuse the two features, in fact, the neural network can better complete the fusion of these two features.

4 Reference

[1].How to use confidence scores in machine learning models,Jonathan Grandperrin, [online],Jan 20, 2021

[2].How to Train a Model: Binary Classifiers, dcolarusso[online],Nov 9,2020