华南理工大学

South China University of Technology

The Experiment Report of *Machine Learning*

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Da Ren, Tao Chu and Huanjie Chen

Supervisor:
Qingyao Wu

Student ID:
201720144900, 201710106550 and
201720145020

Grade:
Graduate

December 21, 2017

Face Classification Based on AdaBoost Algorithm

***Abstract*—The experiment is to complete the face classification using the AdaBoost algorithm. The datasets used contains 500 face and 500 non-face images. The performance of the AdaBoost classifier gets better as the number of base classifiers increases when not over-fitting. The best model we trained use 10 base classifiers and the precision, recall and f1-score of it are all up to 0.93.**

I. Introduction

THIS experiment uses AdaBoost classifier to classify face and non-face from an image dataset with 500 face and 500 non-face images. In the experiment we can understand AdaBoost further and get familiar with the basic method of face detection. Through combining the theory with the actual project, we can experience the complete process of machine learning. We expect that after adjusting the parameters reasonably, the model we have trained in the training set can distinguish between face and non-face images accurately in the verification set.

II. Methods and Theory

AdaBoost is an iterative algorithm. The main idea of AdaBoost is to train a different classifier, i.e., base classifier, in the same training set every iteration. Make the wrong predictive samples more important, and handle it in next iteration. Finally combine these base classifiers to construct a stronger final classifier. A completely AdaBoost algorithm is as follows.

For data sets $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_N)\}$, where $\boldsymbol{x}_i \in X$, $y_i \in \{-1, +1\}$, X is a feature matrix with shape $(sample\_num, feature\_num)$. Every sample has a weight w_i, where $i = 1, 2, ..., N$. In the first iteration initialize $w_i = \frac{1}{N}$. And then enter the iteration. The number of iterations is based on the number of classifiers.

In every iteration, train a base classifier with the sample weights first.

$$h_m(X) = L(D, \boldsymbol{w}_m) \tag{1}$$

where $m = 1, 2, ..., M$. M is the number of base classifiers, i.e, the number of iterations. L is the base classifier. Here we use *sklearn.tree* library *DecisionTreeClassifier* as the base classifier. And the max_depth of the *DecisionTreeClassifier* is set as 2.

And then calculate the error rate of this base classifier.

$$\epsilon_m = \sum_{i=1}^{N} \boldsymbol{w}_m(i) \amalg (h_m(\boldsymbol{x}_i) \neq y_i) \tag{2}$$

If the $\epsilon_m > 0.5$, which means the performance of the base classifier is weaker than random classification, then enter the next iteration directly. Else calculate the weight α_m of this base classifier.

$$\alpha_m = \frac{1}{2} \log \frac{1 - \epsilon_m}{\epsilon_m} \tag{3}$$

Next use the α_m to update the $\boldsymbol{w}_m$

$$\boldsymbol{w}_{m+1}(i) = \frac{\boldsymbol{w}_m(i)}{z_m} e^{-\alpha_m y_i h_m(\boldsymbol{x}_i)} \tag{4}$$

where $i = 1, 2, ..., N$, and

$$z_m = \boldsymbol{w}_m(i) e^{-\alpha_m y_i h_m(\boldsymbol{x}_i)} \tag{5}$$

Repeat the iteration for M times. Finally combine these base classifiers to the final AdaBoost classifier.

$$H(X) = \sum_{m=1}^{M} \alpha_m h_m(X) \tag{6}$$

III. Experiments

A. Dataset

The experiment uses 1000 images, of which 500 are human face RGB images, the other 500 is a non-face RGB images. All the images have been provided by the teacher. We disrupt the dataset randomly. And then select the first 80 percent of the dataset as training set, and the others as validation set.

B. Implementation

After reading the dataset, we first extract the NDP feature of each image and obtain a feature matrix, which including all the images and their features. The model initialization mainly contains 3 parameters, as shown in Table **??**.

TABLE I
Initialization Parameters

Positive sample number	500
Negative sample number	500
Base classifiers number	$2, 5, 10$

And then, use the model described in section II to train an AdaBoost classifier. About the number of base classifiers, we have try $\{2, 5, 10\}$. In order to show the performance improvement of AdaBoost and base classifier, we show the precision, recall and f1-score of the first base classifier and AdaBoost classifiers with 2, 5, and 10 base classifiers at Fig. **??** to Fig. **??**. We can see that the performance of AdaBoost classifier is better than the first base classifier except Fig. 1. It may be because too little base classifiers to use. When there are enough base classifiers, the performance of AdaBoost classifier can obtain a big improvement, as shown at Fig. 3.

In order to intuitively indicate the number of base classifiers on AdaBoost performance, we show the precision, recall and f1-score of AdaBoost classifiers with different number of base classifiers at Fig. **??**. We can see that all of the evaluation

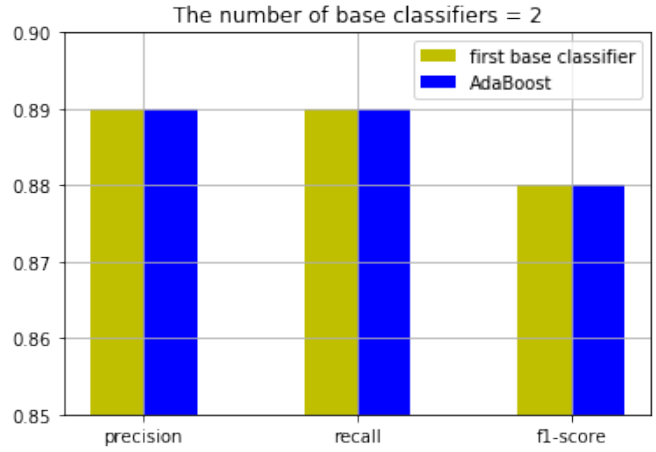

Fig. 1. The precision, recall and f1-score of the first base classifier and AdaBoost classifiers with 2 base classifiers.

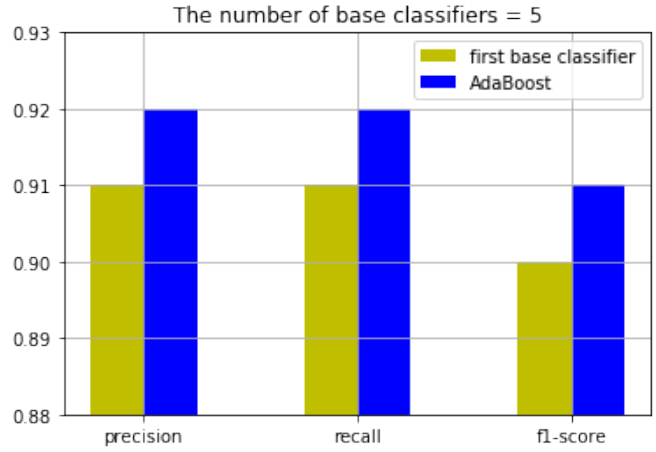

Fig. 2. The precision, recall and f1-score of the first base classifier and AdaBoost classifiers with 5 base classifiers.

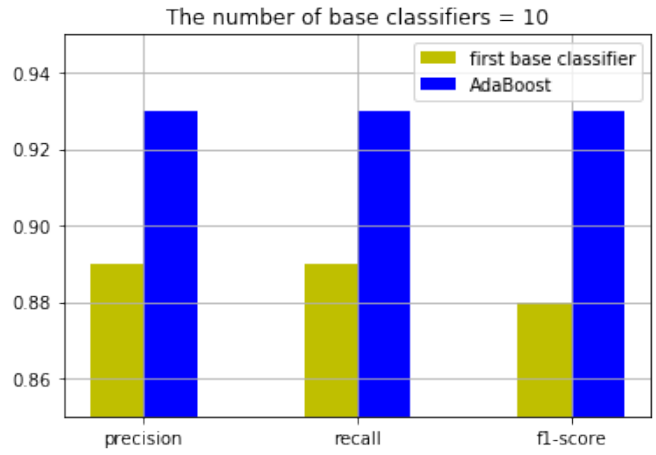

Fig. 3. The precision, recall and f1-score of the first base classifier and AdaBoost classifiers with 10 base classifiers.

metrics get better as the number of base classifiers increases. We don't do the experiment of a very big number of base classifiers, because it will cost much time. But we can easily think that the performance will not get better as the number of base classifiers increases endlessly. There will be over-fitting when the number of base classifiers up to a certain value.

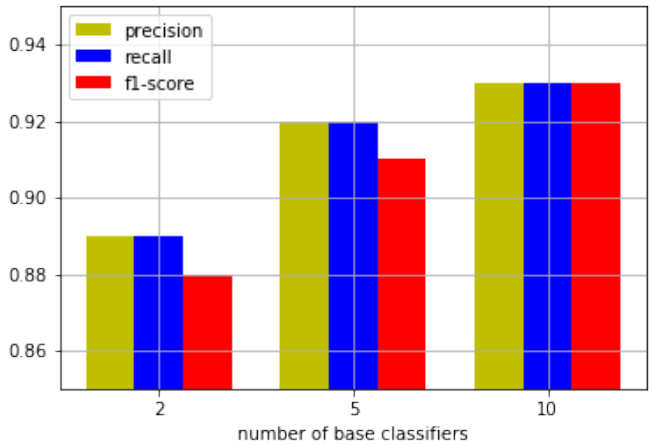

Fig. 4. The precision, recall and f1-score of AdaBoost classifiers with different number of base classifiers.

IV. Conclusion

The experiment complete the whole process of face classification by using AdaBoost algorithm. We analyze and derive the entire AdaBoost algorithm flow. And then we use this algorithm to train and verify 500 face images and 500 non-faces images. We find that the performance gets better as the number of base classifiers increases when there are no over-fitting. In our best model of 10 base classifiers, the precision, recall and f1-score of the model are all up to 0.93.

In the process of experiment, we get further understanding of AdaBoost and machine learning. And because this experiment form is completed in group, our group members collaborate and communicate with each other, solving various problems in the experiment more smoothly than the first two experiments.