華南理工大學

South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Mengdan Zheng, Shoubin Li
and Jing Liang

Supervisor:
Mingkui Tan

Student ID：
201530613795, 201530612040
and 201530741368

Grade:
Undergraduate

December 22, 2017

Face Classification Based on AdaBoost Algorithm

Abstract—In order to understand Adaboost further, get familiar with the basic method of face detection, learn to use Adaboost to solve the face classification problem, and combine the theory with the actual project, we did this experiment, using AdaBoost methods.

I. INTRODUCTION

In this experiment, we should do pre-processing on the image data to extract the NPD feature. Then use the AdaBoostClassifier to train on the train set. Finally do a classification on the validation data to predict whether it is a image with a human face. We evaluate the classifier by the accuracy.

II. METHODS AND THEORY

.

AdaBoost Algorithm:
Boosting, an important integrated learning algorithm, enhances the weak learner to a strong learner with high prediction accuracy. AdaBoost is an acronym for "Adaptive Boosting" in English. In AdaBoost, the samples of the previous basic classifier are warped, and the weighted whole sample is again used to train the next basic classifier. At the same time, a new weak classifier is added in each round until a small enough error rate or a maximum number of iterations is reached.
Step 1, the weight distribution of training data is initialized. Each training sample is given the same weight initially: 1 / N

$$D_1 = (w_{11}, w_{12} \cdots w_{1i} \cdots, w_{1N}), \quad w_{1i} = \frac{1}{N}, \quad i = 1,2,\cdots,N$$

Step 2, Perform multiple rounds of iterations, using m = 1,2, ..., M to indicate the number of iterations:
a. Using a training set with weight distribution Dm, we get the basic classifier:

$$G_m(x): \quad \chi \to \{-1,+1\}$$

b. Calculate the classification error rate of Gm (x) on the training set:

$$e_m = P(G_m(x_i) \neq y_i) = \sum_{i=1}^{N} w_{mi} I(G_m(x_i) \neq y_i)$$

We can know that the error rate, em, of Gm (x) on the training dataset is the sum of the weights of samples misclassified by Gm (x).
c. Calculate the coefficient of Gm (x), where alpha_m represents the importance of Gm (x) in the final classifier.In this step we get the weight of the basic classifier in the final classifier:

$$\alpha_m = \frac{1}{2}\log\frac{1-e_m}{e_m}$$

From the above formula, it can be seen that when em <= 1/2, alpha_m> = 0, and alpha_m increases with the decrease of em, meaning that the smaller the classification error rate, the classifier's role in the final classifier is bigger
d. Update the distribution of weights for the training data set for the next iteration.

$$D_{m+1} = (w_{m+1,1}, w_{m+1,2} \cdots w_{m+1,i} \cdots, w_{m+1,N}),$$

$$w_{m+1,i} = \frac{w_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)), \quad i = 1,2,\cdots,N$$

$$Z_m = \sum_{i=1}^{N} w_{mi} \exp(-\alpha_m y_i G_m(x_i))$$

This makes the weights of the samples misclassified by the basic classifier Gm (x) increase, while the weight of the samples correctly classified decreases. So, in this way, the AdaBoost approach can focus on the more difficult samples.
Step 3. Combine each weak classifier:

$$f(x) = \sum_{m=1}^{M} \alpha_m G_m(x)$$

The final classifier is :

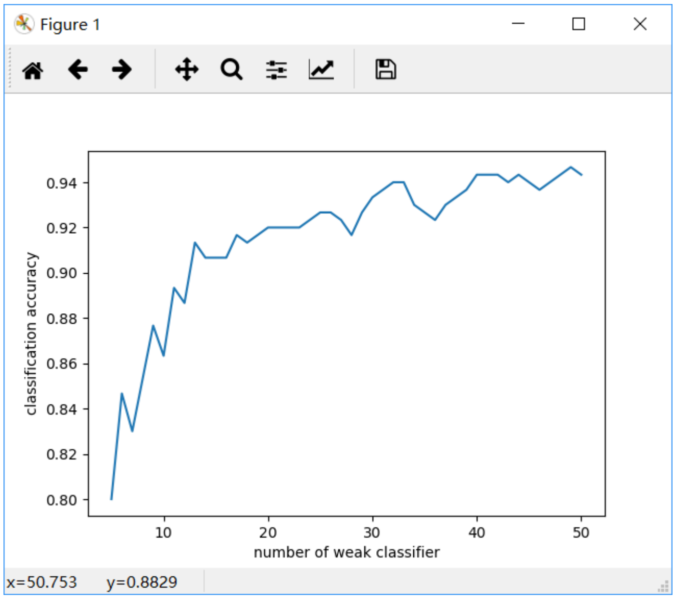
$$G(x) = sign(f(x)) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

III. EXPERIMENT

In this experiment, first we do pre-processing on the image data. We converted the images into a size of 24 * 24 grayscale, get 500 positive samples and 500 negative samples. The data set label is 1 and -1 means with face and without face.Then we using the NPDFeature class in feature.py toextract NPD features. We divided The data set into training set with 700 samples and calidation set with 300 samples randomly.

Then we write the AdaboostClassifier functions based on the reserved interface in ensemble.py and use it to do the face classification. We Initialize each training sample's weight to 1/700. Then train the weak classifier and validation on the validation set. In this step we want to know whether the accuracy will increase as the number of weak classifiers increase. So we let the number of weak classifiers be from 5 to 50 with a fixed max_depth of 4.

Finally, we get the curve graph:



From this chart we can get that the predict accuracy on the validation set increases as the number of weak classifiers increase in general. The best accuracy is about 0.95. We used classification_report () to write predicted result to report.txt.

```
             precision    recall  f1-score   support

         -1       0.96      0.93      0.95       151
          1       0.93      0.96      0.95       149

avg / total       0.95      0.95      0.95       300
```

IV. CONCLUSION

From this experiment, It can be concluded that AdaBoost algorithm has a good effect on face classification . What’s more, the predict accuracy will increase as the number of weak classifiers increase.