The Experiment Report of Machine Learning



Student ID：

201720144894|201721045763|201710106598

Author:

ChangxiZhu|YuzhongYe| YangyangZhao

December 21, 2017

Grade:

 Graduate

Supervisor:

 Qingyao Wu

**SUBJECT:**SOFTWARE ENGINEERING

**SCHOOL:** SCHOOL OF SOFTWARE ENGINEERING

# Face Classification Based on AdaBoost Algorithm

**Abstract**

This paper describes the classification of humans based on Adabooost algorithm. Adabooost have some confusion about is why each iteration to wrong point weight greaten, perhaps by mistake can make the back of the classifier weights are higher, and finish the experiment found our guess is very close to but not entirely, by error of full-time, the next time a classifier after points wrong these points, will enhance the overall error rate, thus led to a smaller, eventually led to the classifier in the hybrid classifier weights of the land. That is to say, this algorithm makes good classifiers account for a higher weight, while the wrong classifier weights are lower.

# **INTRODUCTION**

In order to understand Adaboost further and get familiar with the basic method of face detection, more importantly, It is useful for us to use Adaboost to solve the face classification problem, and combine the theory with the actual project. We experience the project “Face Classification Based on AdaBoost Algorithm” to feel the complete process of machine learning. Furthermore, we use 1000 pictures as the dataset, of which 500 are human face RGF images, the other 500 is a non-face RGB images, which included in the example repository. We experience the project in PyCharm, including following python package sklearn, numpy, matplotlib, pickle, PIL.

# **METHODS AND THEORY**

According to the methods provided by teacher Wu (Steps 1-5 below) and we do it.

1. Read data set data. The images are supposed to converted into a size of 24 \* 24 grayscale, the number and the proportion of the positive and negative samples is not limited, the data set label is not limited.

|  |  |  |
| --- | --- | --- |
| Funtion | Name | import\_data |
| Explanation | details | Data import, judge for file or directory |
| Param | Name | the file or directory name entered |
| Convert type | the type that needs to be converted, which defaults to grayscale |
| Resize | the size that you want to give back is None |
| Return | return an Image object if it is a file, otherwise it returns an Image list, if both are not | |

1. Processing data set data to extract NPD features. Extract features using the NPDFeature class in feature.py. (Tip: Because the time of the pretreatment is relatively long, it can be pretreated with pickle function library dump () save the data in the cache, then may be used load () function reads the characteristic data from cache.)

|  |  |  |
| --- | --- | --- |
| Funtion | Name | PDFeature |
| Explanation | details | It is a tool class to extract the NPD features. |
| image | A two-dimension ndarray indicating grayscale image | |
| n\_pixels | An integer indicating the number of image total pixels | |
| features | A one-dimension ndarray to store the extracted NPD features | |

1. The data set is divisded into training set and calidation set, this experiment does not divide the test set.

|  |  |  |
| --- | --- | --- |
| Funtion | Name | split\_data |
| Explanation | details | The shard data sets are training sets and test sets |
| Param | Data | data set that needs to be syncopated, list type |
| Label | The corresponding label |
| Train\_size | the size of the training set is divided, the default is 0.8, and the default test set is 0.2 |
| Return | the data set after segmentation | |

4.Write all AdaboostClassifier functions based on the reserved interface in ensemble.py. The following is the guide of fit function in the AdaboostClassifier class:

4.1 Initialize training set weights , each training sample is given the same weight.

4.2 Training a base classifier , which can be sklearn.tree library DecisionTreeClassifier (note that the training time you need to pass the weight  as a parameter).

4.3 Calculate the classification error rate of the base classifier on the training set.

4.4 Calculate the parameter according to the classification error rate .

4.5 Update training set weights .

4.6 Repeat steps 4.2-4.6 above for iteration, the number of iterations is based on the number of classifiers.

5.Predict and verify the accuracy on the validation set using the method in AdaboostClassifier.

**Adaboost:**

AdaBoost is an iterative algorithm, and its core idea is different for the same training set training classifier, namely the weak classifier, then put these weak classifier together, to build a stronger final classifier.(many bloggers say that the three cobblers are better than zhuge liang.) The algorithm itself is the realization of data distribution, change it according to the classification of each sample of each training set is correct, the overall classification accuracy as well as the last time, to determine the weights of each sample.The new data for the right value is given to the lower-class classifier for training, and then the classifiers that are received each time are integrated as the final decision classifier.

# Experiment

We did the experiment with classifier 2, 5, 6, 7, 10, and listed the results of 2, 5, 10 below.

**Table 1-1**, the result of one base learner

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| label | precision | recall | f1-score | support |
| -1 | 0.82 | 0.81 | 0.82 | 96 |
| 1 | 0.83 | 0.84 | 0.83 | 104 |
| average | 0.82 | 0.82 | 0.82 | 200 |

**Table 1-2**, the result of two base learner

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| label | precision | recall | f1-score | support |
| -1 | 0.82 | 0.81 | 0.82 | 96 |
| 1 | 0.83 | 0.84 | 0.83 | 104 |
| average | 0.82 | 0.82 | 0.82 | 200 |

**Table 1-3**, the result of five base learner

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| label | precision | recall | f1-score | support |
| -1 | 0.93 | 0.93 | 0.93 | 110 |
| 1 | 0.91 | 0.91 | 0.91 | 90 |
| average | 0.92 | 0.92 | 0.92 | 200 |

**Table 1-4**, the result of six base learner

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| label | precision | recall | f1-score | support |
| -1 | 0.88 | 0.95 | 0.91 | 98 |
| 1 | 0.95 | 0.87 | 0.91 | 102 |
| average | 0.91 | 0.91 | 0.91 | 200 |

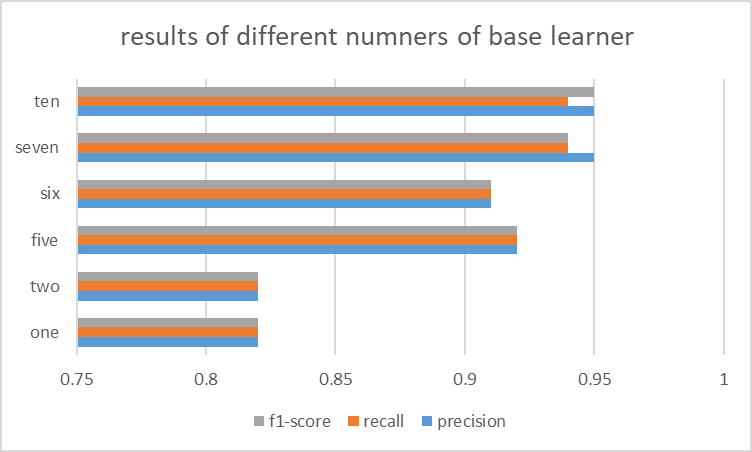
**Table 1-5**, the result of seven base learner

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| label | precision | recall | f1-score | support |
| -1 | 0.95 | 0.92 | 0.93 | 86 |
| 1 | 0.94 | 0.96 | 0.95 | 114 |
| average | 0.95 | 0.94 | 0.94 | 200 |

**Table 1-6**, the result of ten base learner

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| label | precision | recall | f1-score | support |
| -1 | 0.94 | 0.95 | 0.94 | 97 |
| 1 | 0.95 | 0.94 | 0.95 | 103 |
| average | 0.95 | 0.94 | 0.95 | 200 |

We can find that with the growth of numbers of base learner, the average result will be well.



**Fig 1,** results of different numbers of base learner

We also find that with odd numbers of base learner, the results will be better than even numbers of base learner, it seems that the numbers of base learner will be less effected by the threshold, because it’s a problem that when the predict value is just the same as the threshold.

# conclusion

Adabooost have some confusion about is why each iteration to wrong point weight greaten, perhaps by mistake can make the back of the classifier weights are higher, and finish the experiment found our guess is very close to but not entirely, by error of full-time, the next time a classifier after points wrong these points, will enhance the overall error rate, thus led to a smaller, eventually led to the classifier in the hybrid classifier weights of the land. That is to say, this algorithm makes good classifiers account for a higher weight, while the wrong classifier weights are lower.Finally, we can summarize some of the actual scenarios that can be used by the Adaboost algorithm:1) application scenarios for binary classification or multi-classification2) feature selection3) framework used for Boosting bad case correctionBecause the Adaboost algorithm is a simple and easy to implement algorithm. Adaboost algorithm is obtained by combination of weak classifier and strong classifier, both upper and lower bounds of the classification error rate with the increase of training and steady decline, does not have the nature of the fitting and so on, should say is a kind of very suitable for scenarios used in a variety of classification algorithm.