CV Assignment8: Imbalanced Learning

Hao Ding

Due Date: June 30, 2017

1 Introdution

In this assignment, you will implement three approaches for imbalanced learning and compare the performance of different imbalanced learning algorithms by using specific two evaluation criteria.

The whole framework of the implementation for classifying imbalanced dataset is shown in Figure 1, it may serve as a reference for your assignment.

To get started with the assignment, you will need to download the "Multiple Features Data Set" from the website then classify the dataset by three imbalanced learning algorithms. You can change the imbalanced rate of the dataset and evaluate them. At last, your mission is to compare the results which different methods implement with different imbalanced rates.

The details of this assignment are given below in the following sections.

2 Imbalanced learning

In this section, you will implement at least three imbalanced learning algorithms on the dataset to do the classification.

2.1 Methods

- EasyEnsemble[1]

  `http://lamda.nju.edu.cn/code_EasyEnsemble.ashx?AspxAutoDetectCookieSupport=1`

- SMOTE[2]

  `https://cn.mathworks.com/matlabcentral/fileexchange/37311-smoteboost?`

[1] Liu X Y, Wu J, Zhou Z H. Exploratory Under-Sampling for Class-Imbalance Learning. International Conference on Data Mining. IEEE Computer Society, 2006:965-969.

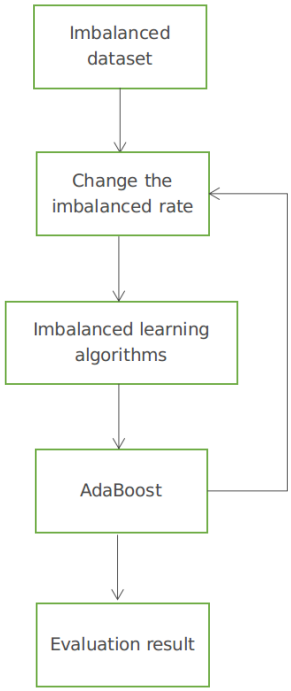
[2] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.

- AdaCost[3]

  `http://read.pudn.com/downloads158/sourcecode/java/javascript/704298/AdaCost.java__.htm`

You need to download the code from the websites which have been given below. The provided ones are already combined with AdaBoost so you do not have to add the algorithm yourself.

Consider that the code of the third assignment use JAVA language,you can consider AdaCost as an optional method.



2.2 Dataset

The dataset you are supposed to use is Multiple Features Data Set for Arabic Sentiment Analysis, which is a balance dataset contains 2000 samples from two attributions in total.

`http://archive.ics.uci.edu/ml/datasets/Multiple+Features` On the website, you can download 6 files with different features. It is enough for you to

[3]Fan W, Stolfo S J, Zhang J, et al. AdaCost: Misclassification Cost-Sensitive Boosting. Sixteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1999:97–105.

use features from one of the files to train your classifier. Except for mfeat-mor, the other files are available for the assignment.

Now the dataset contains 10 classes labled 0 to 9 and each of them has 200 samples. To implement our dichotomy methods, we can categorize the 10 classes into 2 classes. For example, you can regard class 0 to 4 as a class and 5 to 9 as another. Thus the imbalance ratio is 1 which means this is a balance dataset. In the assignment, you need to change this ratio by differently categorize the dataset. The ratio would be between 1 and 9.

3 Evaluation of classification

In this part of the assignment, you will compare the quality of imbalanced learning algorithms using the specific two evaluation measures.

Use such two evaluation measures:

- F-measure = $\frac{2pr}{p+r}$
- G-means = $\sqrt{acc_+ * acc_-}$

Addition: You can also use the AUC[4] to evaluate the quality of imbalanced learning algorithms.

4 Submission

1. Your code
2. A report with your results of explanation and analysis

Zip all your files and submit your assignment to ouceecv@163.com with the subject: YourName_Assignment8.zip. The name of your zip file should be the same as the email subject.

[4] P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997:1145–1159