Introduction to Machine Learning Experiment 2
Ensemble Learning

April 21, 2017

This assignment asks you to implement the ensemble learning algorithms and test your implementation on the given data set.

1 Data

The dataset used in our experiment is WEBSPAM-UK2006[1], which is obtained from a crawler of the .uk domain. There are total 1,803 spam hosts and 4,411 normal hosts which have both content and transformed link features. Every host has 96 content features and 138 transformed link features. 138 transformed link-based features can be divided into five categories: Degree-related features, PageRank-related features, TrustRank-related features, Truncated PageRank-related features and Supporter-related features (The feature file is "./exp2_webspamdata/ContentNewLinkAllSample.csv").

In the experiment, you need to split the data set into training set and test set (4:1, or use cross validation).

2 Tasks

Compare different ensemble learning algorithms with different base classifiers. Two ensemble learning algorithms are required (Bagging and AdaBoost.M1); and two base classifiers are required (SVM and Decision Tree). Thus, you should at least compare 4 combinations: Bagging + DTree, Bagging + SVM, AdaBoost.M1 + DTree, AdaBoost.M1 + SVM.

You are allowed to use existing classifier implementations in the experiment, *but you need to implement the ensemble learning algorithms by yourself.*

3 Optional Tasks

- Try other base classifier (such as K-NN, Naive Bayes...).

[1]Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., Vigna, S.: A reference collection for web spam detection. ACM SIGIR Forum 40(2), 11C24 (2006)

- Analyze the effect of different (kinds of) features (such as using only content features or only link fetures, or doing feature selection).
- Consider feature normalization to improve the performance.
- The data set is in-balanced, how to overcome this.
- Tune the parameters of ensemble learning algorithms, and analyse their effect on performance.

4 Submission

Source code

With necessary comments. No restriction on programming languages, but make sure that TA can run your code easily.

README

A text file that briefly describes how to run your code and produce the reported results. Please also have your name, your student number and your contact information included in it.

Report

A pdf file that includes the following information:

- Your experimental design.
- The experimental results: the results of 4 required combinations. You should report the accuracy metric and you are welcome to use more evaluation metrics or design further experiment to access the performance.
- Your analysis and discussion: For example: why do the algorithms mentioned above perform differently or similarly on the dataset? What is the difference between Bagging and AdaBoost? Which combination is the best one and why?

5 Deadline & Other Information

DEADLINE: Sunday May 7, 2017

Upload the packed file (ZIP format is preferred) with your name and student number in filename to learn.tsinghua.edu.cn. Late submissions WILL NOT BE ACCEPTED[1].

Feel free to contact the TA for further information.

frankyzf94@gmail.com 18671829106

[1] In fact, late submission (through delayed submission channel) will be accepted, but the final score will be multiplied by 0.8

6 Some Existing SVM and DTree Implementations

6.1 SVM

- LibSVM: http://www.csie.ntu.edu.tw/c̃jlin/libsvm/
- SVM-light: http://svmlight.joachims.org/

6.2 DTree

- C4.5: http://www.rulequest.com/Personal/
- C5.0: http://www.rulequest.com/see5-info.html

6.3 Other classifiers

- Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/
- scikit-learn: machine learning in Python http://scikit-learn.org/stable/
- Matlab also has lots of packages for machine learning.

Please note that even if the package provides ensemble learning tools, you SHOULD NOT use them. The implementation of the ensemble learning algorithms (Bagging and AdaBoost.M1) must be done by yourself.