CS 374 Assignment #9
Ensemble Methods

Due the week of April 26, 2021

—

As you discovered last week, computational learning theory makes it possible for us to analyze learning algorithms. It provides tools that allow us to answer questions such as "what is learnable", "how many examples do we need in order to learn a classifier with certain accuracy", etc. In addition, theoretical work often leads to the development of very practical algorithms, such as AdaBoost. This week you will continue your study of AdaBoost, as well as other *ensemble* methods.

Notes and advice as you prepare to do this assignment. This assignment involves an empirical comparison of the performance of several learning algorithms on several sets of data. I recommend that you run the algorithms sooner rather than later. That way, you will have the results available when you're ready to write your report and prepare your presentation for the tutorial session.

What to turn in. For this assignment, you will need to perform an empirical study and write a report. The report can be done either individually or with your tutorial partner.

What to expect during the meeting. Your job during the meeting will be to formally present the material in your report. Whether you use slides, handouts, or the board (virtual or otherwise) is up to you, but I will expect a presentation of your results and analysis, with time for questions and discussion. Because the report is a follow-up to Dietterich's work, you should include appropriate references to background material from that paper in your presentation. You should also be prepared to discuss the paper separately from the presentation itself. Allocate approximately 45 minutes for your presentation and 30 minutes for the paper discussion.

1 Boosting and Bagging

1.1 Reading

Please read the following:

- Alpaydin, Sections 18.1-18.4 (4th edition) or 17.1-17.4 (3rd edition)
- Alpaydin, Sections 18.6-18.7 (4th edition) or 17.6-17.7 (3rd edition)
- Since the reading in Alpaydin refers to the bias-variance tradeoff, you might also find it useful to read Sections 4.3 and 4.7 (both editions).

2 The Effect of Noise

We have often discussed the potential impact of noise on classifier learning. This week you will have the opportunity to explore the effects of class noise on ensembles of decision trees and other learners.

2.1 Reading

Please read "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization" by Dietterich, which appeared in the Machine Learning Journal in 2000. In this paper, Dietterich compares the effectiveness of randomization, bagging, and boosting for improving the classification accuracy of C4.5 on 33 data sets from the UC Irvine Machine Learning Repository. He finds that AdaBoost performs better, on average, than Bagging. However, when class noise is added to the data sets, he finds that, on average, Bagging outperforms AdaBoost. This has been observed by others as well. (Why might we expect AdaBoost to be affected more significantly by noise in the data?)

2.2 An Experimental Comparison of Two Methods for Constructing Ensembles: Bagging and Boosting

The results presented in Dietterich's paper are quite extensive and convincing. However, one of the issues we've discussed frequently this semester is reproducibility. This week you will empirically assess boosted and bagged decision trees, both with and without class noise.

In addition, it is interesting to consider whether we would observe the same general behavior with a different base classifier. So you will also explore the behavior of Boosting and Bagging with other learners as your base classifiers: Naive Bayes and 3-NN.

The product of your investigation will be a summary of your findings.

2.2.1 Empirical Evaluation: Comparing Algorithms on Data Sets with No Noise

Before assessing the impact of class noise on learning, it is important to determine the baseline performance of each algorithm:

- J48 (with pruning turned off)
- AdaBoostM1 with J48 as base classifier (with pruning turned on)
- Bagging with J48 as base classifier (with pruning turned off)

as well as the same for Naive Bayes and 3-NN. You already know how to find J48, Naive Bayes, and 3-NN in Weka. You'll find AdaBoostM1 and Bagging in the "meta" folder. You will need to set some parameters for AdaBoostM1 and Bagging. For AdaBoostM1, set the number of iterations to 50. You can also modify the seed for the random number generator, if you'd like. Don't modify "useResampling", and don't worry about the weight threshold. For Bagging, set the number of iterations to 100 and the bagSizePercent to 100. Again, you can modify the seed for the random number generator. Be sure to set the base classifier appropriately in all cases. That includes setting the parameters for J48. For instance, be sure that pruning is turned off for Bagging, but be sure it's on for AdaBoost. That is, for AdaBoost set the "unpruned" parameter to "False" and the confidence factor to 0.10.

The data sets to be evaluated are as follows:

- vehicle
- sick
- kr-vs-kp
- audiology
- hypothyroid

You can find the data sets in

```
/home/faculty/andrea/shared/cs374/EnsembleData
```

as well as on Glow. You'll note that all of these are were tested by Dietterich. In his experiments, Dietterich only pruned the basic C4.5 decision trees for one of the five sets, so we will turn off pruning for J48. He did not prune the bagged trees at all for these data sets, so we will turn off pruning of the underlying J48 trees for bagging. He pruned the boosted trees for four of the five data sets, which is why we will turn on pruning for AdaBoost. This won't replicate the experiments exactly, but it will be close.

For all of the above, perform 10 x 10-fold cross validation. (This is the default setting in the Weka Experimenter.) Your report should give the error percentage of each algorithm on each of the data sets. You can refer to Table 1 in Dietterich's paper for ideas on how to present the results.

2.2.2 Empirical Evaluation: Comparing Algorithms on Data Sets with Noise

Now you're ready to assess the impact of noise on each of the algorithms. I've already prepared the necessary files for you. You'll find noisy versions of each of the above data sets in the same EnsembleData directory. There are three noise levels per data set: 5%, 10%, and 20%.

Now run each of the algorithms again (doing 10 x 10-fold cross-validation), this time on the noisy data sets. (First compare the 5% noise sets. Then the 10%. And finally the 20%.) Give tables with the resulting error percentages. You can, again, follow the format suggested in Table 1 of the paper. How do your results compare to Dietterich's? Are you observing the same general trends?

Don't forget that you need to consider J48 (alone, bagged, and boosted), Naive Bayes (alone, bagged, and boosted), and 3-NN (alone, bagged, and boosted).

In addition to providing the tables and giving a high level summary of the highlights, list and briefly discuss four interesting observations or other points that are relevant to the experiments you just completed.

2.3 The Tutorial Meeting

For the tutorial meeting, be prepared to:

- Explain the differences between bagging and boosting.
- Present your empirical results and observations, making appropriate references to the Dietterich paper.
- Discuss the Dietterich paper.