Big Data Analytic Programming Report

Wang Bo
S0204742 21/11/2016

1. Learning curve comparison

Figure 1 and figure2 shows how the three types of classifier's accuracy varies as a function of the number of training examples. For the first half of training data, VFDT very quickly achieve about 95% accuracy. Naïve Bayes and logistic regression however, has worse accuracy and did not evolve with the number of training samples. For the second half of data, the VFDT accuracy dropped very fast during the data phase changing, however, with number of samples increasing, the VFDT accuracy arise quickly and converge with the accuracy of the logistic regression and naïve Bayes.

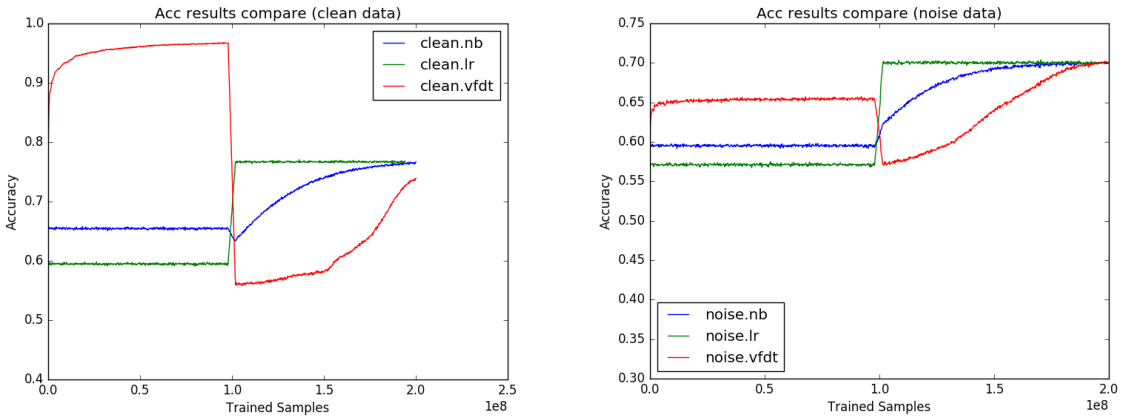

Figure 1: Training results comparison for naïve Bayes(nb), logistic regression(lr) and very fast decision tree(vfdt), apply clean dataset(left) and noise dataset(right).

2. Experiments

2.1 Logistic regression.

Research questions: what is the effect on the training speed and accuracy by changing the lambda and ETA parameters?

Figure3 show 5 different learning curves with different values of parameters as indicated on the figure. For curves 2,3,4,5, the training did not finished due to the limited time. However, it can be show that the change of lambda and ETA do not have too much influence on the accuracy. The learning speed, however, due to the unfinished training, cannot be explored in this figure.

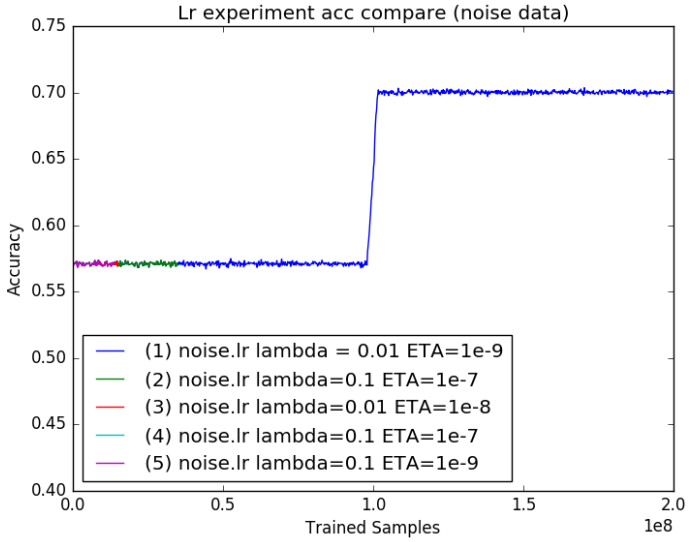

Figure 3: Training results comparison for logistic regression, with difference training parameters (lambda: penalty coefficient, ETA: learning rate).

2.2 Very fast decision tree.

Research questions: what is the effect on the training speed and accuracy by changing the Tau and Nmin parameter in VFDT?

Figure4 show 4 learning curves with different values of parameters as indicated on the figure. Again, due to limited time, curve (2) and (3) did not finished. However, by compare curve (3) and curve (1), it can be shown that by decreasing of Tau, the accuracy at the beginning of training is lower, this might because we made very strict rule for splitting ties with smaller Tau, which made the tree grow more rigidly.

By changing the Nmin from 200 to 20000, the trees have different learning rate for the last half of data.

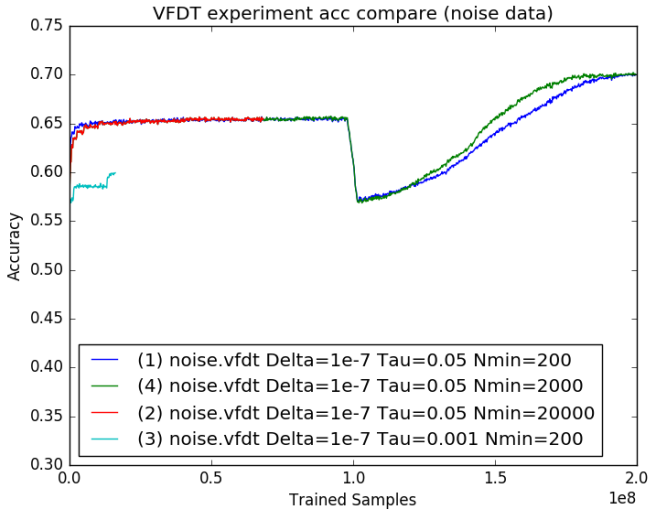

Figure 4: Training results comparison for VFDT, with difference training parameters of learning with the default parameters.