In the real world, data generally tends to be noisy, missing and inconsistent due to their huge size and their possible multiple, heterogenous sources, which leads to low quality data mining results. Therefore, analysts sometimes would preprocess data, such as data transformation, data cleaning, data integration, data reduction, etc. In this section, we would experiment whether the data transformation, the discretization and the binarization process, could reduce the bias and improve the confidence interval coverage for S-Learner, T-Learner and X-Learner using random forests (RF) and bayesian additive regression trees (BART).

Discretization

We used histogram to categorize continuous values into discrete integers, which depended on the number of the bins in the histogram. If more bins were used in the histogram, the original floating-point values would be represented more accurate in the resulting discrete dataset.

Binarization

We used a unique technique, binary trees, to convert continuous values into binary digits. These digits indicated the intervals where the continuous values fell into. Each branching was decided by either the mean or median of the subsets within that interval. In addition, the height of the binary trees would decide the accuracy and the information loss of the resulting binary dataset.

Bias Reduction

We reproduced the bias and RMSE experiments on the GOTV data. All S-Learner, T-Learner and X-Learner used RFs and BARTs on the original floating-point data, the discrete integer data and the binary bit data. Due to the computation power, limited memory size and training time, we only estimated the biases and RMSEs from 1000 training size and 50 testing size up to 20000 training size and 1000 testing size. The result is in the figure.

From the figure, we could see that the discrete data and the binary data did decrease the RMSE, because the discretization and binarization could reduce the noises. The noise reduction would be beneficial for models which were sensitive towards noises, like linear models. However, the non-linear RF and BART would not take this advantage from the discrete data and the binary data, for these two models are robust and not traditionally sensitive to noises. The discretization and binarization would only increase the biases, since each feature of these datasets covers a range of values instead of one accurate floating-point value.

Confidence Interval (CI) Coverage

As the author mentioned in their paper, the estimation of the bootstrap confidence interval requires massive computation power and huge memory consumptions. Due to the limited resources of our machine, we could not fully replicate the original full confidence interval simulation with exactly the same setting of the original paper. Our experiments used 10,000 training examples and 500 testing examples running 100 bootstrap sampling to estimate confidence interval. The CI coverage experiment results are shown in the figure.

Because previously we had learned the discretization and binarization would not reduce the biases, we could expect that the discrete and binary dataset would not exceed the original floating-point dataset and achieve the ideal 95% CI coverage. The CI coverage experiment proved our suspection. From the results, we could see that none of the S-Learners, T-Learners and X-Learners and their given datasets achieved the target. Also, the linearity of the confidence interval coverage and the average confidence interval length matched the results from the experiments of the original paper.

As a result, the discretization and binarization did not reduce the bias and improve RMSE score, which was probably because the learners use the non-linear models, the RF and BART, that the feature transformation preprocess did not have huge impact on these models. Thus, the learners using these two models had already achieved their state-of-the-art performance. On the other hand, the bootstrap sampling still suffered the bias from the CATE estimators thus could not produce the 95% CI coverage. Therefore, the discretization and binarization would not adjust the bias, improve the learners using RF and BART and provide the target CI coverage.