wang-jialu-assignment4

Jialu Wang (jwang282@illinois.edu)

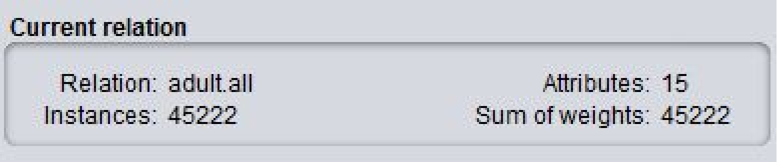
1. Data processing

Before test with weka, we need to deal with the two train and test files. The process is as follows:

- change the extension of files to turn the files into .csv files.
- add header to both files
- remove unknown values (instances with “?”)
- change the test data. ‘>/<=50K.’ in test file to ’>/<=50K’

2. Data description

After processing, there are 30162 instances in train data and 15060 instances in the test data. For convenience and compatibility purposes, we copy and paste the test data below the train data file so there are 45222 instance in total (provided as the supplementary file ‘ adult.all.csv)



Current relation

Relation: adult.all

Instances: 45222

Attributes: 15

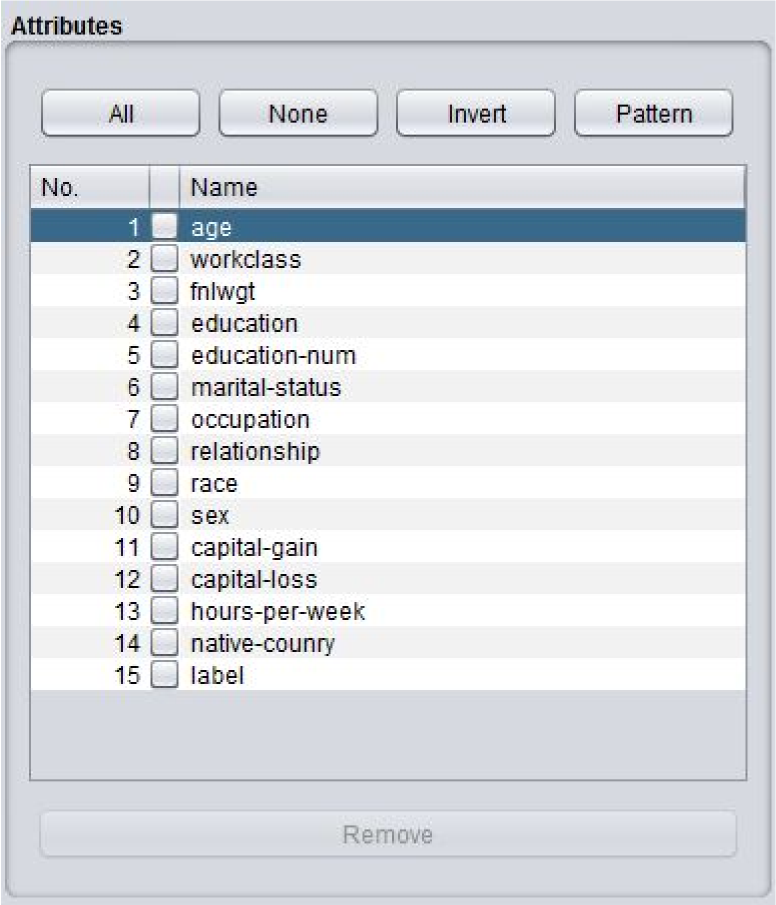
Sum of weights: 45222

The 14 attributes and possible values are:

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary,

Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

And the lable: <=50K or >50K



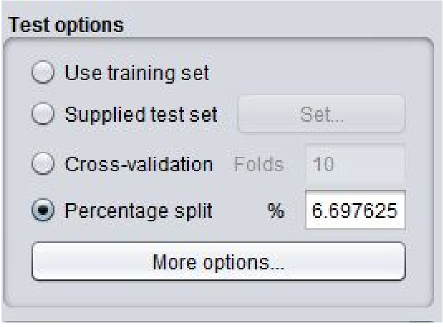
No.		Name
1	□	age
2	□	workclass
3	□	fnlwgt
4	□	education
5	□	education-num
6	□	marital-status
7	□	occupation
8	□	relationship
9	□	race
10	□	sex
11	□	capital-gain
12	□	capital-loss
13	□	hours-per-week
14	□	native-counry
15	□	label

3. Test with classifiers

In order to replicate the methods used in the paper and use the Bayes Net, we choose 3 classifiers as follows:

- weka.classifiers.bayes.BayesNet
- weka.classifiers.trees.J48 (confidence level 0.5) (same as C4.5)
- weka.classifiers.trees.NBTree

As we put the train and test data in a single file. So in the test option, we choose percentage split with (30162/45222=) 66.697625%.



○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

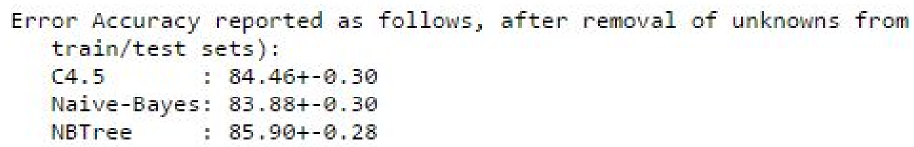
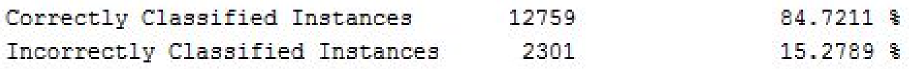
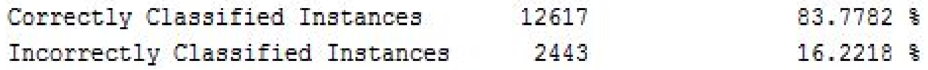
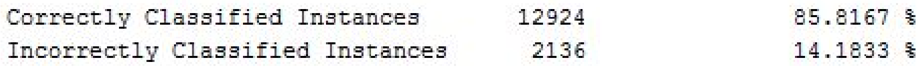
● Percentage split % 6.697625

4. Results

As the full replicated results from Weka 3.8.1 are too long in content, they are provided in the supplementary files as ‘BayesNet’,’J48’ and ‘NBTree’.

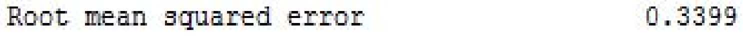
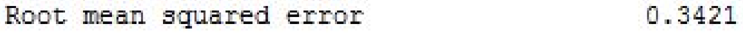
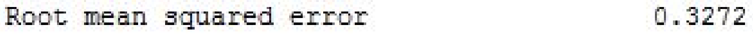
- If we compare the accuracy rate with the result of paper

The results of the paper is obtained from the source: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names

Paper Results	 Error Accuracy reported as follows, after removal of unknowns from train/test sets): C4.5 : 84.46+-0.30 Naive-Bayes: 83.88+-0.30 NBTree : 85.90+-0.28
J48	 Correctly Classified Instances 12759 84.7211 % Incorrectly Classified Instances 2301 15.2789 %
BayesNet	 Correctly Classified Instances 12617 83.7782 % Incorrectly Classified Instances 2443 16.2218 %
NBTree	 Correctly Classified Instances 12924 85.8167 % Incorrectly Classified Instances 2136 14.1833 %

Replicated results are all within the range provided by the paper results
The classifier NBTree has the largest accuracy rate.

- If we compare the RMSE:

J48	 Root mean squared error 0.3399
BayesNet	 Root mean squared error 0.3421
NBTree	 Root mean squared error 0.3272

The classifier NBTree has the smallest error.

- If we compare the decision trees:

As by the results of methods provided in supplementary files, we can also find that NBTree has far fewer nodes than J48.

5. Conclusion

After data cleaning, we reach the same instance amounts as with the paper. Our replicated results are all within the range of those provided by the paper. Generally speaking, from the perspective of accuracy rate, error terms and the complexity of decision trees, the classifier NBTree performs better in the study and estimation of the dataset adult.all.