1 Cleaning and Preprocessing

1.1 Data Cleaning

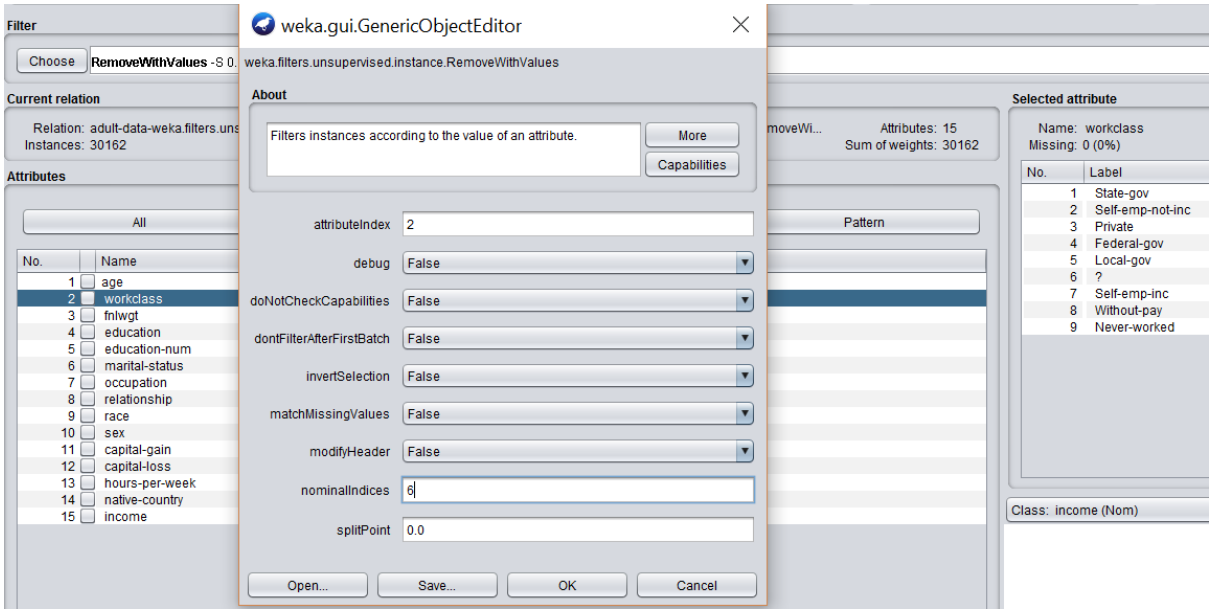
Neither the training set “adult.data” nor the test set “adult.test” has headers of attributes and class, so I manually add the 14 attribute names and one class label called “income” to the first line of the two datasets. In “adult.test”, an extra period (compared with “adult.data”) occurs at the end of each line, so I delete all of them before loading into Weka. In this way, for the class column, it can only have two values of “<=50K” or “>50K” (no period).

Besides, the training set and test set do not have csv file extension, but actually the data are delimitated by comma. Therefore, I consider directly change the file type of the two datasets to csv file, and then import them to Weka.

In addition, an important thing I find is that the data of each row in “adult.data” are actually separated by comma followed by a space. While in “adult.test”, the delimitators are not consistent, comma or comma and a space. When I delete the white space, the question mark inside the data cannot be identified when imported to Weka. So, I modify the training and test sets in format separated by comma and a space. Finally, I save the two files as csv files and import them to Weka.

1.2 Remove Unknown Values

After importing into Weka, I remove the unknown values with label “?” for the nominal data. For attribute “workclass”, “occupation” and “native-country”, I remove all the instances which have value of “?” as label. I choose the filter called RemoveWithValues to do the work, as is shown below.



Similarly, I remove all the “?” instances in the test set.

Consequently, I save the two cleaned datasets as arff files, called “adult-data.arff” for training set and “adult-test.arff” for test set.

2 Introduction of Datasets

2.1 General Statistics

There are 45222 rows in all if instances with unknown values are removed (train=30162, test=15060).

Duplicate or conflicting instances: 6

Probability for the label ">50K": 24.78% (without unknowns)

Probability for the label "<=50K": 75.22% (without unknowns)

Prediction task is to determine whether a person makes over 50K a year.

2.2 14 Attributes + 1 Class

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt (final weight): 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.

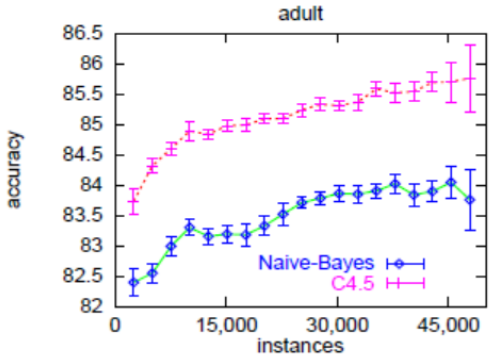
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

2.3 Methods (Classifiers)

- Naïve-Bayes
- Decision Trees (C4.5)
- NBTree (hybrid of Naïve-Bayes and Decision Trees)

2.4 Performance Measures

- Accuracy rates with error bars of 95% confidence intervals



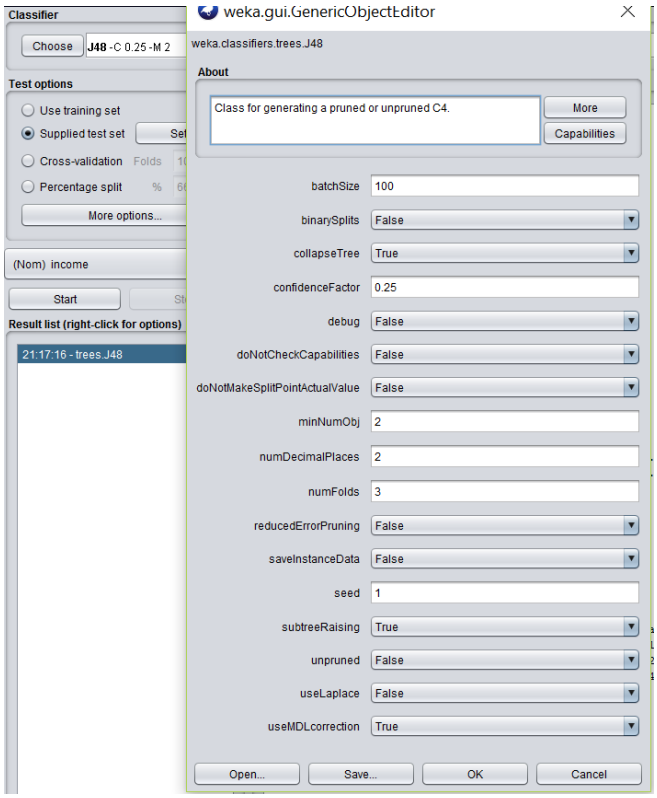
3 Replication

3.1 C4.5 Classifier

Input “adult-data.arff”, and use supplied test set “adult-test.arff”.

Choose Classifiers -> trees -> J48

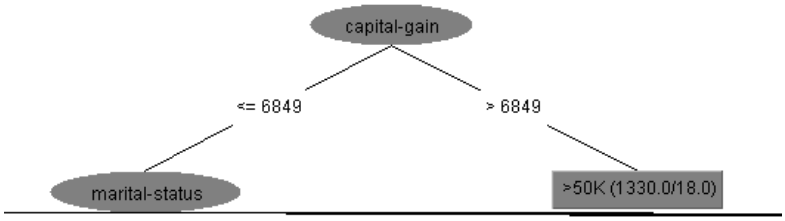
Allow “InputMappedClassifier” for suppled test set




```
Number of Leaves  :     590

Size of the tree :      754

Attribute mappings:

Model attributes                      Incoming attributes
----------------------------          ----------------
(numeric) age                    --> 1 (numeric) age
(nominal)  workclass             --> 2 (nominal)  workclass
(numeric)  fnlwgt                --> 3 (numeric)  fnlwgt
(nominal)  education             --> 4 (nominal)  education
(numeric)  education-num         --> 5 (numeric)  education-num
(nominal)  marital-status        --> 6 (nominal)  marital-status
(nominal)  occupation            --> 7 (nominal)  occupation
(nominal)  relationship          --> 8 (nominal)  relationship
(nominal)  race                  --> 9 (nominal)  race
(nominal)  sex                   --> 10 (nominal)  sex
(numeric)  capital-gain          --> 11 (numeric)  capital-gain
(numeric)  capital-loss          --> 12 (numeric)  capital-loss
(numeric)  hours-per-week        --> 13 (numeric)  hours-per-week
(nominal)  native-country        --> 14 (nominal)  native-country
(nominal)  income                --> 15 (nominal)  income
```

Result and Discussion:

The top of the decision tree (not complete) (not binary relationship):



The **marital status** firstly matters for the income prediction. The second factor is **education-num**.

The accuracy rate and the confusion matrix:

```
=== Summary ===

Correctly Classified Instances       12848               85.3121 %
Incorrectly Classified Instances      2212               14.6879 %
Kappa statistic                          0.5814
Mean absolute error                      0.2005
Root mean squared error                  0.3281
Relative absolute error                 53.8645 %
Root relative squared error             76.2233 %
Total Number of Instances            15060

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.928    0.378    0.883      0.928    0.905      0.585    0.885     0.948     <=50K
               0.622    0.072    0.739      0.622    0.675      0.585    0.885     0.735     >50K
Weighted Avg.  0.853    0.303    0.848      0.853    0.849      0.585    0.885     0.896

=== Confusion Matrix ===

     a     b   <-- classified as
 10547   813 |     a =  <=50K
  1399  2301 |     b =  >50K
```

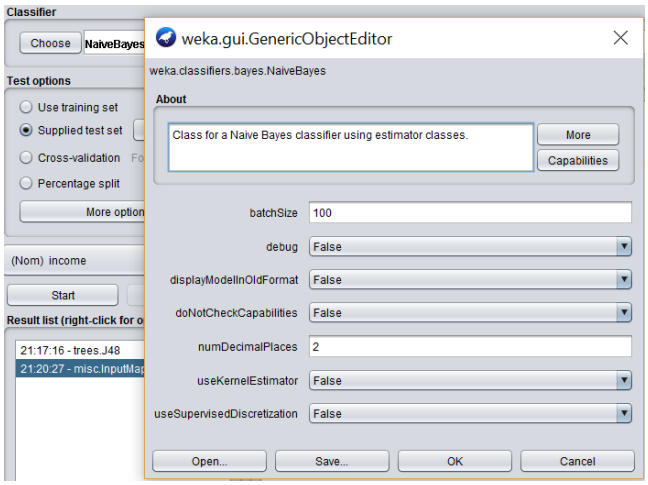
The accuracy rate of C4.5 presented in “adult.names” file is 84.46 ± 0.30 %. The actual accuracy rate for my replication is 85.3121%.

3.2 Naïve-Bayes Classifier

Input “adult-data.arff”, and use supplied test set “adult-test.arff”.

Choose Classifiers -> bayes -> NaiveBayes

Allow “InputMappedClassifier” for suppled test set




```
Attribute mappings:

Model attributes                      Incoming attributes
----------------------------          ----------------
(numeric) age                    --> 1 (numeric) age
(nominal)  workclass             --> 2 (nominal)  workclass
(numeric)  fnlwgt                --> 3 (numeric)  fnlwgt
(nominal)  education             --> 4 (nominal)  education
(numeric)  education-num         --> 5 (numeric)  education-num
(nominal)  marital-status        --> 6 (nominal)  marital-status
(nominal)  occupation            --> 7 (nominal)  occupation
(nominal)  relationship          --> 8 (nominal)  relationship
(nominal)  race                  --> 9 (nominal)  race
(nominal)  sex                   --> 10 (nominal)  sex
(numeric)  capital-gain          --> 11 (numeric)  capital-gain
(numeric)  capital-loss          --> 12 (numeric)  capital-loss
(numeric)  hours-per-week        --> 13 (numeric)  hours-per-week
(nominal)  native-country        --> 14 (nominal)  native-country
(nominal)  income                --> 15 (nominal)  income
```

Result and Discussion:

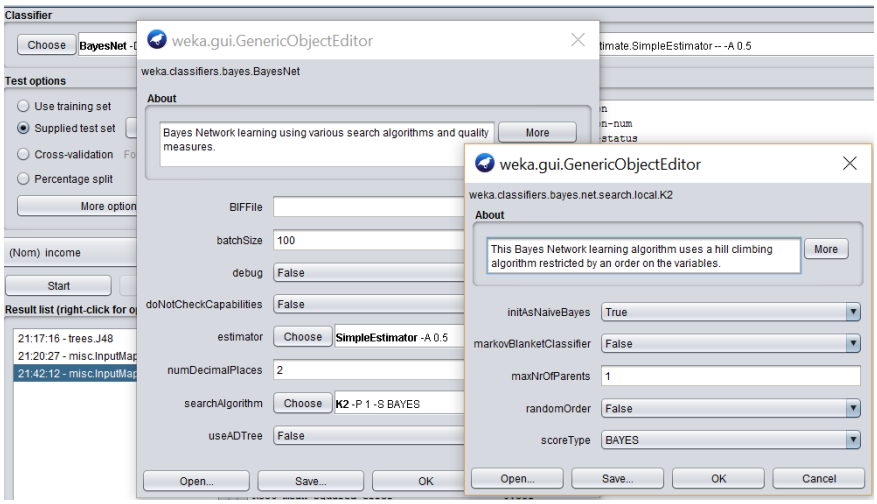
The accuracy rate and the confusion matrix:

```
=== Summary ===

Correctly Classified Instances       12430               82.5365 %
Incorrectly Classified Instances      2630               17.4635 %
Kappa statistic                          0.4811
Mean absolute error                      0.181
Root mean squared error                  0.381
Relative absolute error                 48.6033 %
Root relative squared error             88.5072 %
Total Number of Instances            15060

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.929    0.492    0.853      0.929    0.889      0.491    0.888     0.962     <=50K
               0.508    0.071    0.699      0.508    0.588      0.491    0.888     0.721     >50K
Weighted Avg.  0.825    0.389    0.815      0.825    0.815      0.491    0.888     0.903

=== Confusion Matrix ===

     a     b   <-- classified as
 10550   810 |     a =  <=50K
  1820  1880 |     b =  >50K
```

The accuracy rate of Naïve-Bayes presented in "adult.names" file is 83.88 ± 0.30 %. The actual accuracy rate for my replication is 82.5365%.

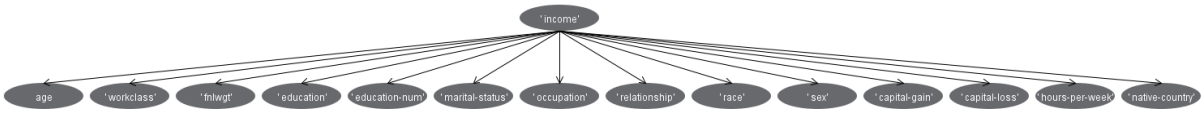
Another method to use Naïve-Bayes:

Choose BayesNet and set maxNumberofParents to 1 for K2 search algorithm. -> Naïve-Bayes



Result and Discussion:

Visualize graph:



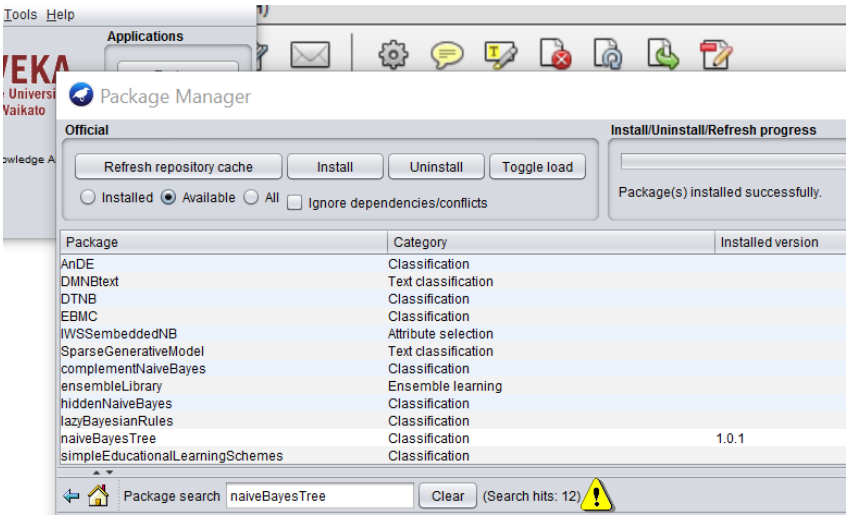
There is no correlation between attributes. Each node only has one parent.

```
=== Summary ===

Correctly Classified Instances       12630               83.8645 %
Incorrectly Classified Instances      2430               16.1355 %
Kappa statistic                          0.5987
Mean absolute error                      0.1783
Root mean squared error                  0.3438
Relative absolute error                 47.8972 %
Root relative squared error             79.87   %
Total Number of Instances            15060

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.852    0.202    0.928      0.852    0.888      0.606    0.916     0.971     <=50K
               0.798    0.148    0.637      0.798    0.708      0.606    0.916     0.796     >50K
Weighted Avg.  0.839    0.189    0.857      0.839    0.844      0.606    0.916     0.928

=== Confusion Matrix ===

    a    b   <-- classified as
 9679 1681 |    a =  <=50K
  749 2951 |    b =  >50K
```

The accuracy rate of Naïve-Bayes in this experiment is 83.8645%.

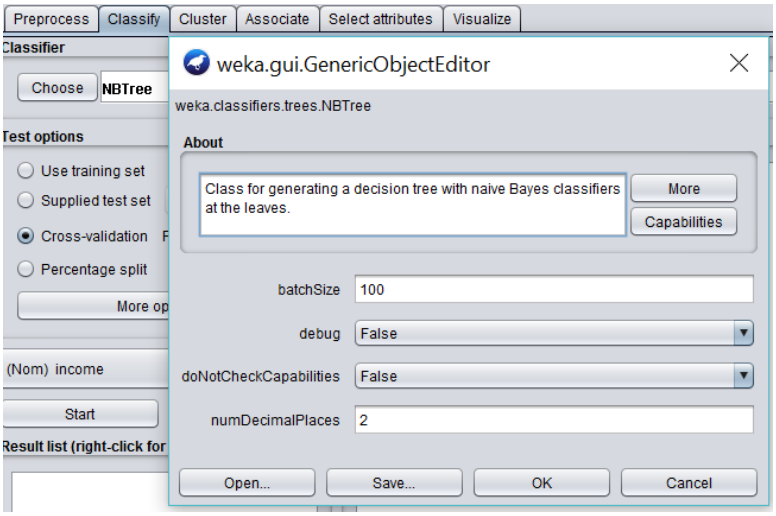
3.3 NB Tree Classifier

Weka -> Tools -> Package Manager -> install “naiveBayesTree”



Input “adult-data.arff”, and use supplied test set “adult-test.arff”.

Choose Classifiers -> trees -> NBTree



Allow “InputMappedClassifier” for suppled test set

Result and Discussion:

```
=== Summary ===

Correctly Classified Instances       12924               85.8167 %
Incorrectly Classified Instances      2136               14.1833 %
Kappa statistic                          0.6014
Mean absolute error                      0.1673
Root mean squared error                  0.3296
Relative absolute error                 44.9399 %
Root relative squared error             76.5565 %
Total Number of Instances            15060

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.925    0.348    0.891      0.925    0.908      0.603    0.905     0.964     <=50K
                 0.652    0.075    0.740      0.652    0.693      0.603    0.905     0.785     >50K
Weighted Avg.    0.858    0.281    0.854      0.858    0.855      0.603    0.905     0.920

=== Confusion Matrix ===

     a     b   <-- classified as
 10511   849 |     a =  <=50K
  1287  2413 |     b =  >50K
```

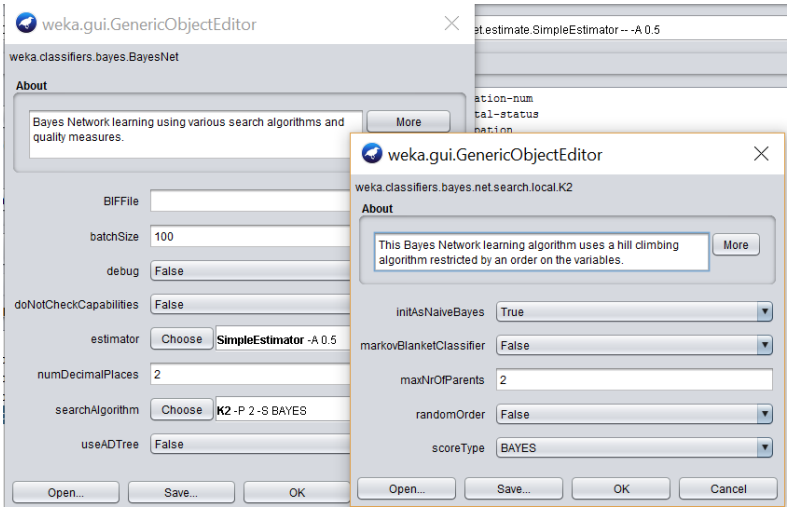
The actual accuracy rate of NBTree is 85.8167%. The given accuracy rate in “adult.names” file is 85.90±0.28 %.

4 Test on Bayesian Net Classifier

4.1 Method 1

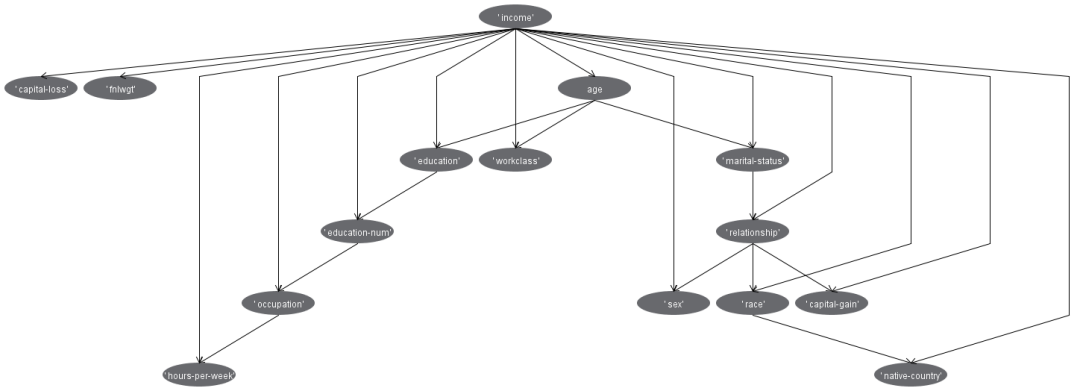
Choose BayesNet and set maxNumberofParents to 2 for K2 search algorithm.

Do not use AD Tree.



Result:

Visualize graph (the number of parents for each node is 1 to 2):




```
=== Summary ===

Correctly Classified Instances       12907               85.7039 %
Incorrectly Classified Instances      2153               14.2961 %
Kappa statistic                          0.6043
Mean absolute error                      0.1859
Root mean squared error                  0.3106
Relative absolute error                 49.9373 %
Root relative squared error             72.1382 %
Total Number of Instances            15060

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.917     0.328     0.896      0.917     0.906      0.605    0.918     0.971      <=50K
               0.672     0.083     0.726      0.672     0.698      0.605    0.918     0.810      >50K
Weighted Avg.  0.857     0.268     0.854      0.857     0.855      0.605    0.918     0.932

=== Confusion Matrix ===

     a     b   <-- classified as
 10422   938 |     a =  <=50K
  1215  2485 |     b =  >50K
```

The accuracy rate of Bayes Net (K2 search algorithm, max 2 parents) is 85.7039%.

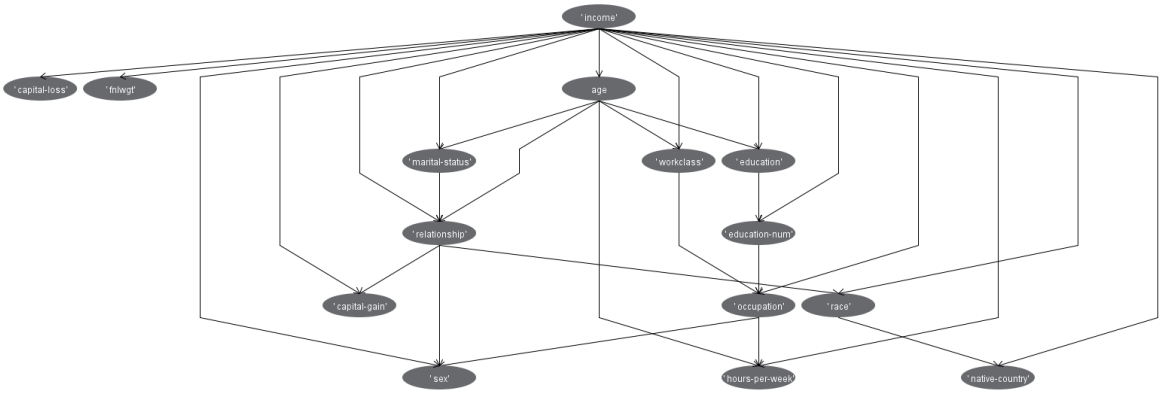
4.2 Method 2

Choose BayesNet and set maxNumberofParents to 3 for K2 search algorithm.

Do not use AD Tree.

Result:

Visualize graph (the max number of parents for each node is 3):



```
=== Summary ===

Correctly Classified Instances       12924               85.8167 %
Incorrectly Classified Instances      2136               14.1833 %
Kappa statistic                          0.602
Mean absolute error                      0.1834
Root mean squared error                  0.3102
Relative absolute error                 49.2554 %
Root relative squared error             72.0603 %
Total Number of Instances            15060

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.925    0.346    0.891      0.925    0.908      0.604    0.918     0.971     <=50K
                 0.654    0.075    0.739      0.654    0.694      0.604    0.918     0.810     >50K
Weighted Avg.    0.858    0.279    0.854      0.858    0.855      0.604    0.918     0.932

=== Confusion Matrix ===

     a     b   <-- classified as
 10503   857 |     a =  <=50K
  1279  2421 |     b =  >50K
```

The accuracy rate of Bayes Net (K2 search algorithm, max 3 parents) is 85.8167%.

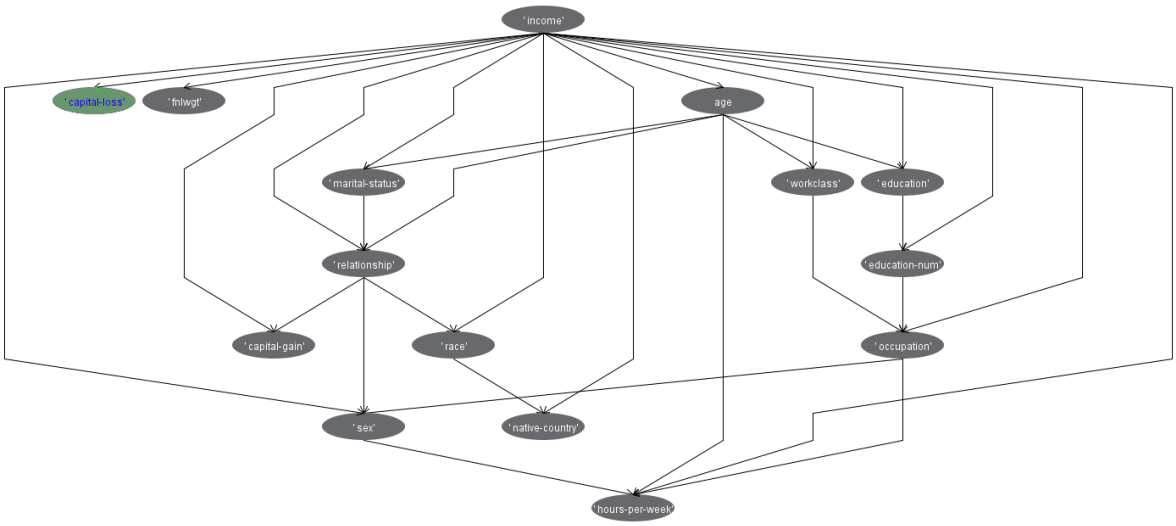
4.3 Method 3

Choose BayesNet and set maxNumberofParents to 4 for K2 search algorithm.

Do not use AD Tree.

Result:

Visualize graph (the max number of parents for each node is 4):



```
=== Summary ===

Correctly Classified Instances       12921               85.7968 %
Incorrectly Classified Instances      2139               14.2032 %
Kappa statistic                          0.6022
Mean absolute error                      0.1837
Root mean squared error                  0.3108
Relative absolute error                 49.3545 %
Root relative squared error             72.1929 %
Total Number of Instances            15060

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.924    0.344    0.892      0.924    0.907      0.604    0.917     0.971     <=50K
               0.656    0.076    0.737      0.656    0.694      0.604    0.917     0.808     >50K
Weighted Avg.  0.858    0.278    0.854      0.858    0.855      0.604    0.917     0.931

=== Confusion Matrix ===

     a     b   <-- classified as
 10492   868 |     a =  <=50K
  1271  2429 |     b =  >50K
```

The accuracy rate of Bayes Net (K2 search algorithm, max 4 parents) is 85.7968%.

4.4 Method 4

Choose BayesNet and set maxNumberofParents to 5 for K2 search algorithm.

Do not use AD Tree.

The result is the same as Method 3. The graph remains the same. The max number of parents for each node in the model is still 4.

4.5 Method 5

Using **AD Tree** for Method 1 to 3.

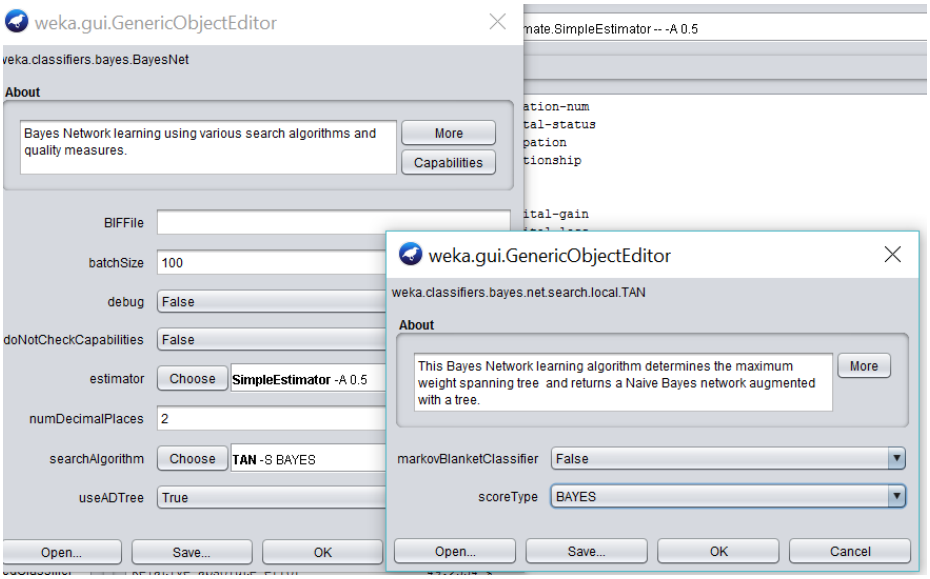
The accuracy rate and the graph remain the same. The time taken to build the model rises, but the time taken to test model on supplied test usually decreases a bit, but not always.

The AD Tree is a very sparse data structure to minimize memory use. It can be used to accelerate Bayes net structure finding algorithms.

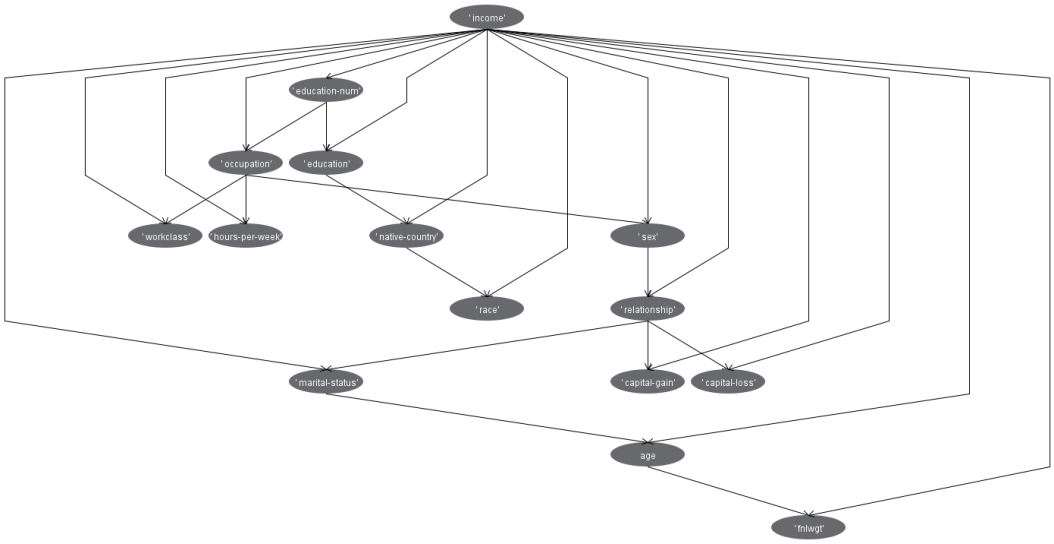
4.6 Method 6

Change search algorithm to TAN (Tree Augmented Naïve Bayes):

Using AD Tree.



Visualize graph:




```
=== Summary ===

Correctly Classified Instances       12895               85.6242 %
Incorrectly Classified Instances      2165               14.3758 %
Kappa statistic                          0.6032
Mean absolute error                      0.1853
Root mean squared error                  0.3127
Relative absolute error                 49.7614 %
Root relative squared error             72.6331 %
Total Number of Instances            15060

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.916    0.326    0.896      0.916    0.906      0.604    0.917     0.971     <=50K
               0.674    0.084    0.722      0.674    0.697      0.604    0.917     0.808     >50K
Weighted Avg.  0.856    0.267    0.853      0.856    0.855      0.604    0.917     0.931

=== Confusion Matrix ===

     a     b   <-- classified as
 10401   959 |     a =  <=50K
  1206  2494 |     b =  >50K
```

The accuracy rate for TAN search algorithm is 85.6242%.

For TAN, it starts with naïve Bayes, and considers adding second parent to each node (apart from class node). For this experiment, it is less effective than Bayes Network K2 search algorithm.

5 Comparison of Different Classifiers

Classifier	Actual Accuracy Rate (%)	Given Accuracy Rate (%)
C4.5 J48	85.3121	84.46±0.30
Naïve-Bayes	82.5365	83.88±0.30
Naïve-Bayes (Bayes-Net K2 max=1) w/ AD Tree	83.8645	83.88±0.30
NBTree	85.8167	85.90±0.28
Bayes-Net K2 max=2 w/ AD Tree	85.7039	
Bayes-Net K2 max=3 w/ AD Tree	85.8167	
Bayes-Net K2 max=4+ w/ AD Tree	85.7968	
Bayes-Net TAN w/ AD Tree	85.6242	

It can be found that Naïve-Bayes performs the worst, and Decision Tree (C4.5) performs better, and then Bayes-Net of all search algorithms performs better than the above two. The NBTree method proposed in the paper also performs the best.

Among the above classifiers, for this domain, Bayes Network with K2 search algorithm, and setting max number of Parents for each node to be three, performs the best accuracy rate, which is 85.8167%. Besides, NBTree classifier also maintains the highest accuracy rate for this dataset.