Ex3: Classification (2)

Tokyo Tech.
Intro. to Comp. & Data
Exercise&hw week3ex

Classification rule discovery project:
- How to evaluate and use obtained models.

1. Homework assignment #3.

please send one pdf file via email to
Suzukakedai: watanabe.o.aa-cd18s@ml.m.titech.ac.jp
Ookayama: watanabe.o.aa-cd18o@ml.m.titech.ac.jp
before week2lect *of each campus*

2. Some explanation on Weka.

Samples used in ex3 can be found in Weka's data directory.

1. Homework assignment #3: Task

Your task #1: Understand statistical values

(a) Use credit-g.arff (given as a sample data in Weka) to study the meaning of stat. data on a obtained decision tree for credit-g.arff.

1. randomly select 70% instances and make a data set c700rnd.arff also make a data set c300left.arff consisting of the remaining set.

2. use J4.8 to create a decision tree from c700rnd.arff.
   * set J4.8's property "minObjnum" as "5", and
   * execute J4.8 under the test set mode "Percentage split" with default 66%, which means to use 66% of c700rnd.arff for creating a decision tree and use the remaining 34% (i.e., 238 instances) for testing.

3. then obtained the decision tree (and its performance on the test data), examine the following points.
   (next page)

1. Homework assignment #3: Task

Your task #1: Understand statistical values

(a) Use credit-g.arff (given as a sample data in Weka) to study the meaning of stat. data ...

...

3. then obtained the decision tree (and its performance on the test data), examine the following points.

+ re-evaluate the obtained tree on c300left.arff (for comparing with the error prob. on the test set).

Ex #2 page5

+ select several "typical" leaf nodes of the obtained tree and check the # of error instances (and its ratio) in c300left.arff that reach these nodes (for comparing with the error prob. on c700rnd.arff).

page 11

page 8

Warning: Unfortunately, it seems that Weka gives this info. only on the whole set (and not on the test set).

Your task #2: Create better and/or useful rules

Consider the data set breast-cancer.arff again. You might have found difficulty for getting a good classifier. Also you might have noticed the unbalance between the two class values, i.e., recurrence-event (let's consider it “negative” here) and no-recurrence-event. Suppose that we would like to reduce false positive rate under this situation. What can we do?

(b) Try to make a rule with relatively small false-positive rate by giving more weight to negative instances.

* Here again use bc114rnd.arff for the training phase and bc172left.arff for the test phase.

demo. in the ex. session

* Use supervised filter "ClassBalancer" or supervise filter "Resample" for creating a training set (from bc114rnd.arff) for emphasizing negative instances.
* Try not only J4.8 but also NaiveBayes for classifier and obtain two or three candidate classification models with different features.
* You can also evaluate the F value on the test set. (Warning: the test set should be used under the original distribution.)

Your task #2: Create better and/or useful rules (Cont.)

(c) Derive "rules" with the true negative rate >(almost) 70%.

no need to be exact

in the week3lecture

page 10

Here by "rule" we mean a decision method including the usage of a model (e.g., a decision tree).

+ obtain at least three rules,
  * you may use those obtained at (b) and the others, but when using the models obtained at (b), be careful not to use the balanced sample set for the testing phase.
+ compare them and determine the most appropriate one, and
+ explain why you think that your chosen rule is the best.

1. Homework assignment #3: Report

submit through OCW *before* week4lect

Required items that you need to explain:

Japanese is OK!!
About 1 page for each item, please!

From Task #1:

(1) Describe the obtained decision tree on c700rnd.arff:

+ the decision tree with error rate info. on each leaf (on the training set), and
+ error rate on the test set, false positive rate, false negative rate.

(2) Explain the result of your experiment (a), and give your comment from what you learned in the week 3 lecture.

(e.g.)

Derive how many test instances are necessary for estimating error probability within ε with 99% confidence, and discuss your results with this bound.

1. Homework assignment #3: Report (Cont.)

Required items that you need to explain:

From Task #2:

(3) Describe the models, i.e., decision tree and naive Bayes obtained at (b). Discuss whether what you proposed as "almost best" in the Ex#2 (i.e., the prev. ex.) is still good or not; explain why you think so.

(4) Describe the obtained rules at (c); explain how to get your best rule (and how to test its performance) and why you think that your rule is best compared with the other rule.

Optional:

(5) Write your questions that you had from this exercise.

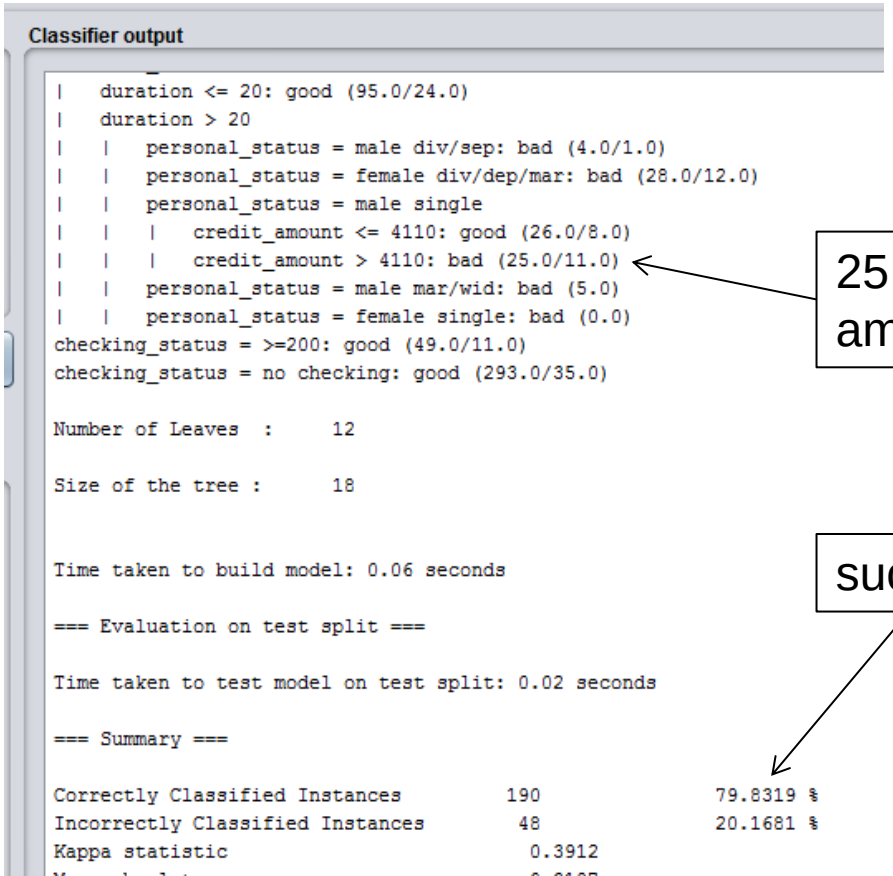
You may get extra credit by asking technically interesting question(s)!

2. Tips for using Weka

demo. in the ex. session

How to see the obtained decision tree:

- After obtaining a decision tree under, e.g., "Percentage and split" with default 66%, Weka reports the obtained decision tree as follows.



the stat. results of the obtained decision tree on the whole (i.e., 700) instances

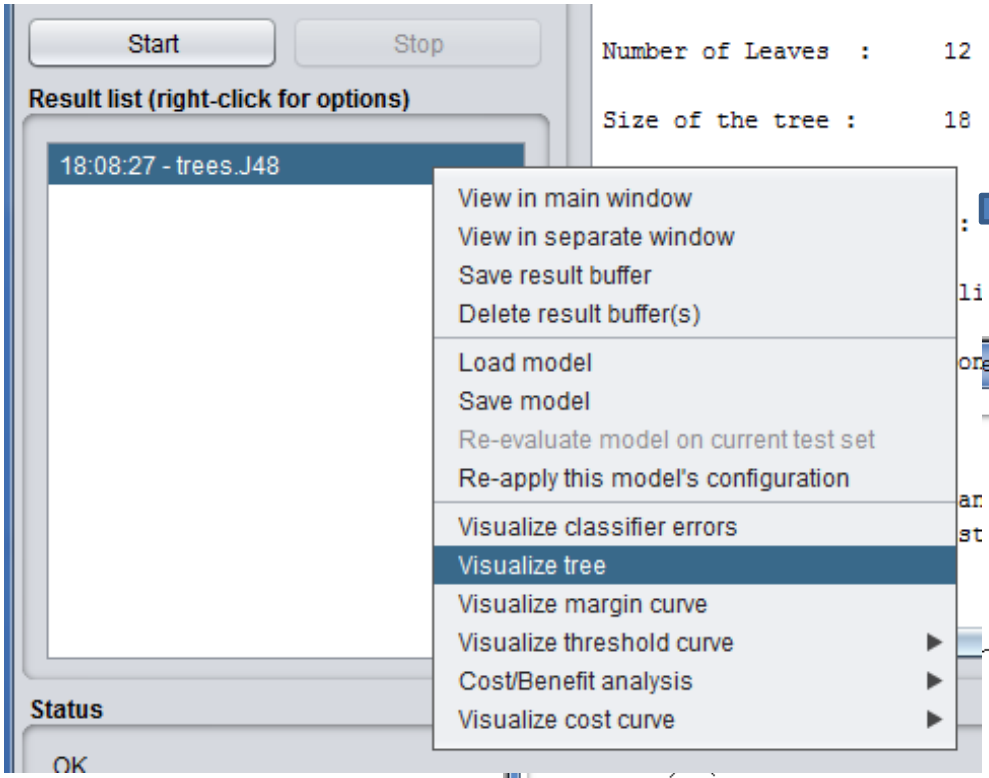
25 instances reach this node,
among which 11 instances are misclassified

success rate (on the 238 test instances)

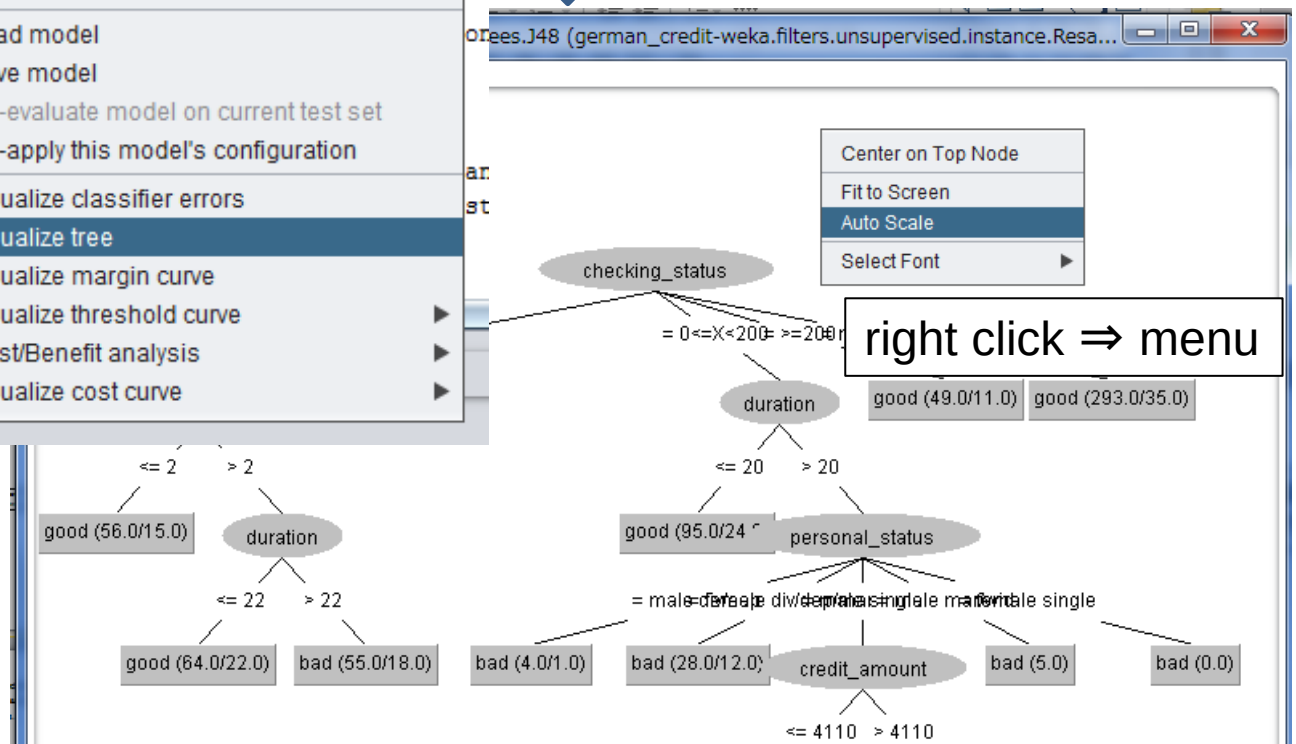
How to see the obtained decision tree:

- You can see the obtained decision tree visually.

  right click ⇒ menu ⇒ "Visualize tree"



displayed tree size can be controlled as follows.

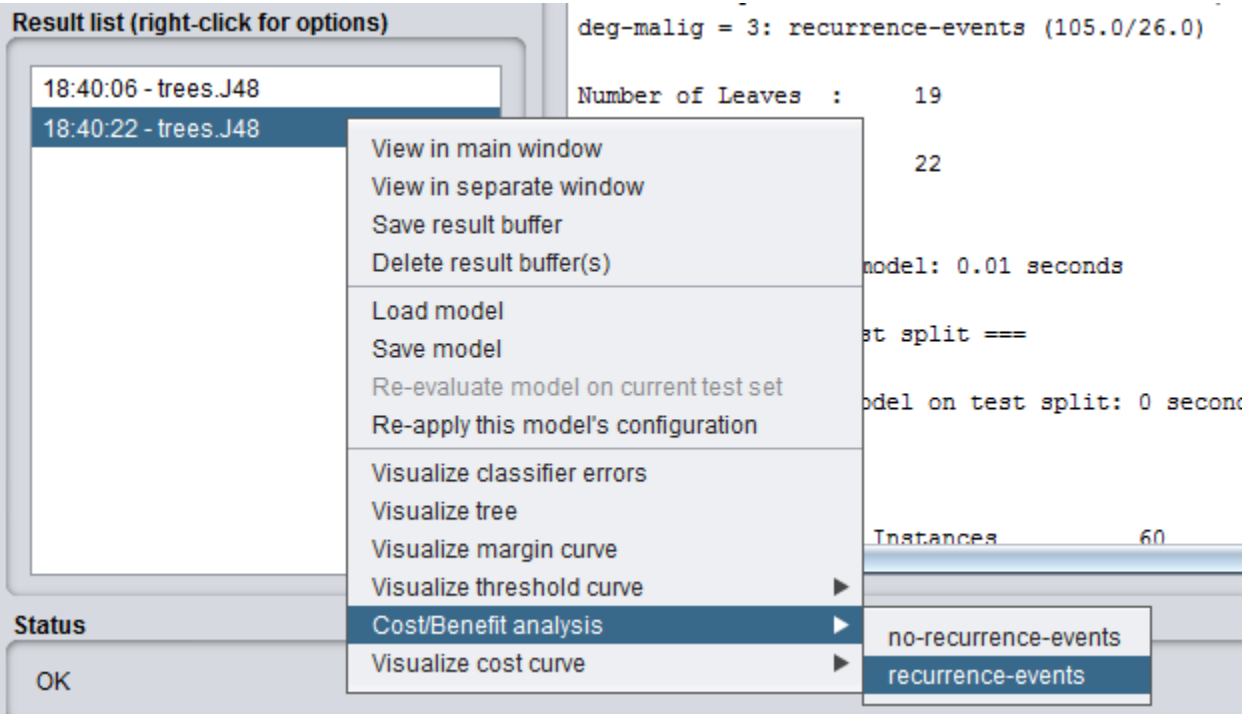


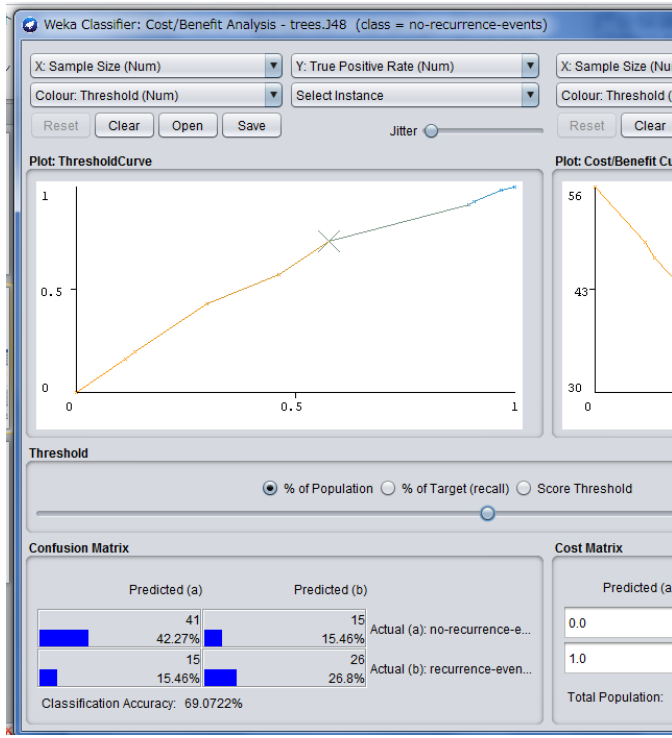
right click ⇒ menu

How to analyze the obtained decision tree:

- For cost/benefit analysis of the obtained model

  right click ⇒ menu ⇒ "Cost/Benefit analysis" & recurrence-events



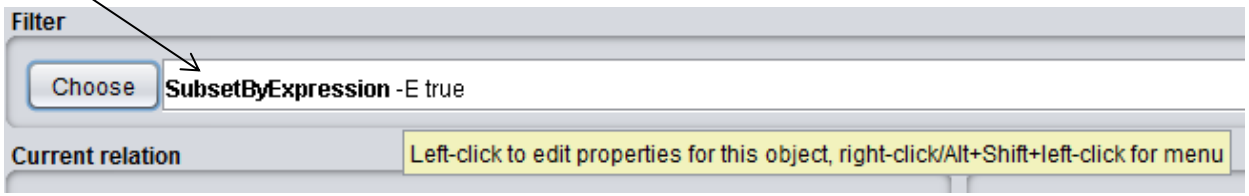


Choosing specific instances:

demo. in the ex. session

- For selecting instances satisfying some specific conditions

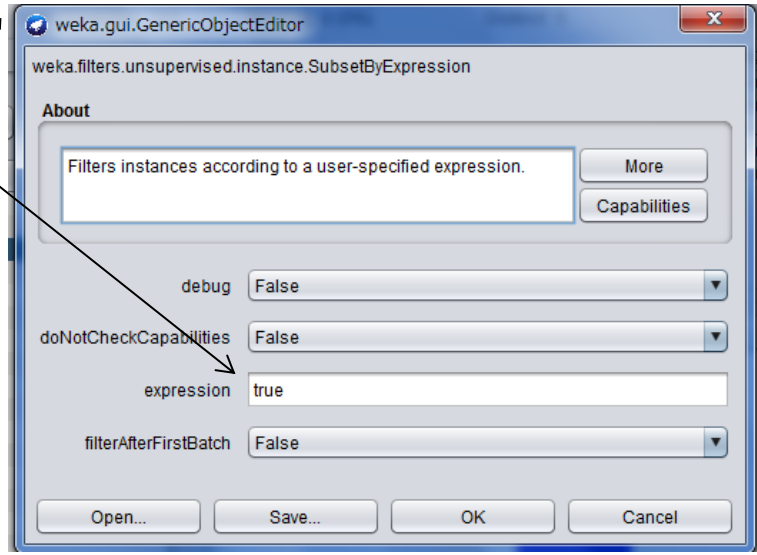
1. "Explorer" ⇒ "Preprocess" ⇒ "Open file ..." to open your target file.
2. "Choose" ⇒ "filters" ⇒ "unsupervised" ⇒ "instances" ⇒ "SubsetByexpression"
3. Click here to open a window for setting parameters.



4. Set a Boolean expression to "expression"
   e.g.,
   (ATT1 = "<0") and (ATT3 ="all paid")

attribute index or attribute no.

attribute value called "Label"



Then "ok" ⇒ "Apply".