Ex2: Classification (1)

Tokyo Tech.
Intro. to Comp. & Data
Exercise&hw week2ex

Classification rule discovery project:
- Introduction to Weka[*1].
- Use 1R, Naive Bayes, and Decision Tree methods.

1. Homework assignment #2.

   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.

*1 Weka is constructed and provided by the University of Waikato

1. Homework assignment #2: Task

Your task #1: Try to use Weka

(a) Prepare m8126.arff. page 8, 9

by, e.g., one of the following methods:

- prepare by yourself from the one used in the ex1.
  m8126rnd.txt + mushroomspec.txt ==> the arff file

- use a file conversion tool of Weka.
  m8126rnd.txt ==> m8126rnd.csv ==> the arff file

(b) Compute classification models on m8126.arff.

A "rule" is called a model in Weka

+ try the following three classification tools:
  OneR, Naive Bayse, and J4.8.
  * use "Percentage split" with default 66%.
+ for J4.8, try several "hard" situations to confirm that this dataset is a relatively easy for the classification task.

page 4

Your task #2: Understand basics on classification tasks and decision trees

(c) Use breast-cancer.arff (given as a sample data in Weka) to learn how to create an appropriate decision tree.

+ use J4.8 and use 40% (114 instances) of the whole set, and obtain *your best decision tree* under this situation.

The difference may be small ;-)

* Conduct your experiment under the following setting:

+ use randomly selected 114 instances for training and use the remaining 172 instances for testing.

page 10

* let us name these sample sets as bc114rnd.arff and bc172left.arff.

+ "Use training set" option on bc114rnd.arff, and use bc172left.arff as a test set by "Supply test set".

page 5

* Compare at least three different decision trees.

page 11

Remarks & Suggestions:

- For the task (b), consider several situations to make the classification task harder and confirm that we can still get a decision tree with good performance. For example, reducing the number of instances, removing important attributes, etc.

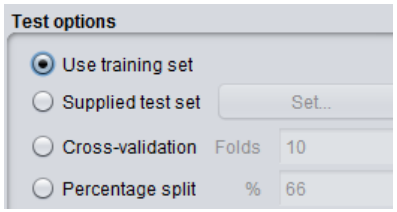
page 9

Remarks & Suggestions:

- For the task (c), do not use "Cross validation" or "Percentage split" because we cannot control obtained decision trees.

demo.
in the ex. session

1.

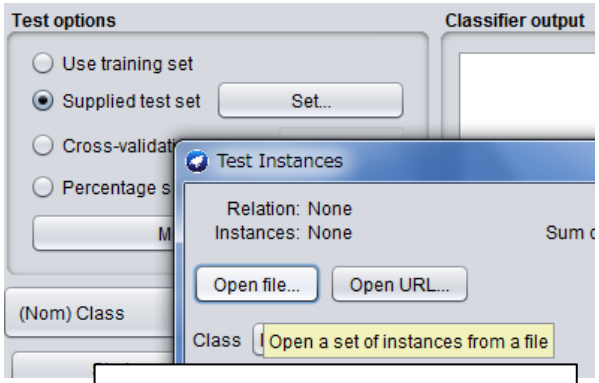


use bc114rnd.arff

Learning Phase

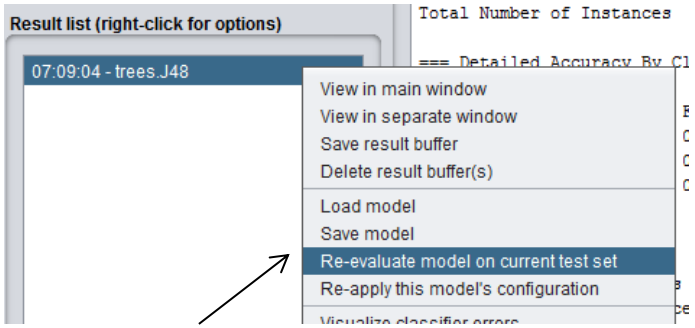
2.

Testing Phase



specify bc172left.arff

3.



use "Re-evaluate"

1. Homework assignment #2: Report

submit through OCW *before* week3lect

Required items that you need to explain:

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

From Task #1:

(1) Describe the obtained three models and the following statistical data on these models:

A "rule" is called a model in Weka

+ the size of training set and test set,
+ accuracy on the training set, and
+ accuracy, true positive rate, and true negative rate on the test set.

(2) Explain what you did to confirm the "easiness" of the mushroom dataset, and explain why you can conclude that the classification task is relatively easy on it.

1. Homework assignment #2: Report (Cont.)

Required items that you need to explain (Cont.):

From Task #2:

(3) Describe the obtained decision tree and the following statistical data:

+ accuracy on the training set,
+ accuracy on the test set, and
+ accuracy, true positive rate, and true negative rate on the test set.

(4) Explain why you think that your obtained decision tree is *best* (or *almost best*) under the situation of Task #2 comparing with the other (at least) two decision trees.

Optional:

(5) Write your questions that you had from Task #2.

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

2. Tips for using Weka

Throughout this course, we will basically use only the "Explorer" application of Weka. For the standard explanation see the following web page:

https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Here we give some additional tips that may yelp you for conducting our exercises.

demo. in the ex. session

Preparing your data: "Explorer" ⇒ "Preprocess"

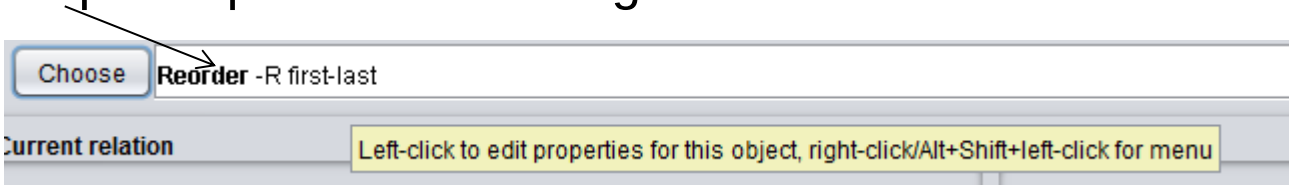
- To open your target file: "Open files ..."
  * Need to go up to, e.g., desktop, to find your folder.
  * Weka supplied data can be found at, e.g.,
    "C:Users/local/Program Files/Weka-3-8/data"
    ↑ this may differ for each PC
  * Specify "all file" mode to find csv files.
  * For preparing a arff file or a csv file, see the above web page for Weka.

Preparing your data: "Explorer" ⇒ "Preprocess"

- After opening your target file, you might want to do the following:

  - reorder attributes (for moving the class attribute which is the 1st attribute in the mushroom data to the last for simplifying your later tasks in Weka):

    (1) "Choose" ⇒ "filters" ⇒ "unsupervised" ⇒ "attributes" ⇒ "Reorder"
    (2) Click here to open a parameter setting window.



    (3) In the param. setting window, set "attributeIndices" as "2-23,1".

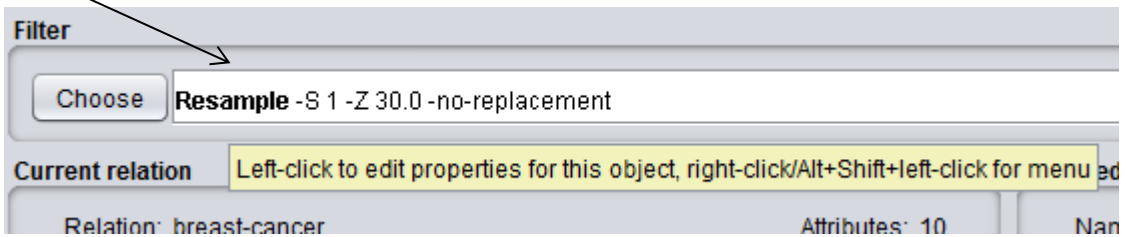
  - remove some attributes (for making harder data mining tasks):

    (1) Simply mark attributes that you want to remove, and then "Remove".

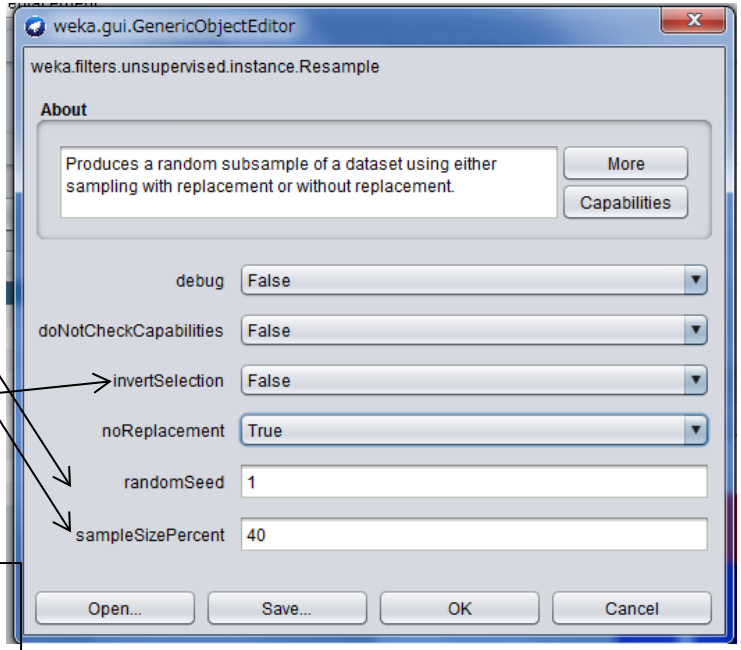
Preparing your data: "Explorer" ⇒ "Preprocess"

- select instances randomly (for making a training set and test set):

(1) "Choose" ⇒ "filters" ⇒ "unsupervised" ⇒ "instances" ⇒ "Resample"
(2) Click here to open a window for setting parameters.



(3) For selecting 40% instances randomly set param.s like this and then "Apply".
(4) After saving the selected instances, "Undo", and then open this window again to change "invertSelection" to true to select remaining 60% instandces.



"invert" means to select the *un*chosen ones.

Classification: "Explorer" ⇒ "Classify"

* of course, after specifying your target dataset.

- For making decision trees: Choose the classifier J4.8.

* For changing obtained trees, click here to open a window for setting J4.8's properties.