SYNOPSIS

data2text – convert feature-vector data to a single text-valued field.

eliminate – eliminate obviously redundant conditions from a ruleset

filter-text – perform feature selection on text, optionally converting it to an attribute-value format.

pair – do a paired test

partition – randomly subsample a dataset

pprint-rules – print a ruleset

predict – makes predictions using a ripper-generated ruleset

rocchio – strawman text categorizer

select-class – remove all but two classes from a dataset

summarize – summarize output of predict.

test-rules – test individual ripper rules on data

verify – check a dataset for consistency

DESCRIPTION

Miscellaneous programs to assist in evaluating the rulesets produced by RIPPER. All of these programs will summarize their legal options if invoked with a -h option. Most of them also will read from standard input if invoked with a -s option.

PREDICT

predict [options] filestem

Uses the ruleset filestem.hyp to predict classes of the examples in the file filestem.test. (If the -s option is used then examples are taken from standard input instead of filestem.test). Output is a series of lines of the form

predicted-class p n actual-class

where p and n are the numbers of positive and negative examples classified by the rule used to pick the predicted-class.

SUMMARIZE

summarize [options] filestem

Summarize the predictions appearing on the standard input. Reports the error rate, recall, precision, and a confusion matrix.

VERIFY

verify [options] filestem

Do a consistency check on the dataset filestem.data

ELIMINATE

eliminate [options] filestem

Eliminate redundant conditions from the ruleset filestem.hyp

TEST-RULES

test-rules [options] filestem

Gives a rule-by-rule breakdown of error rates of the rules in filestem.hyp on the examples in filestem.test

PAIR

pair predfile-1 predfile-2

Does a paired test of the classifiers used to generate the predictions in the files predfile-1 predfile-2. Predfile-1 and predfile-2 should be the outputs of predict. Outputs the following information: number of "wins" (times the classifier used for predfile-1 is right and the classifier of predfile-2 is wrong); number of "losses";

number of "ties" (times both classifiers agree); estimate of the probability of a win given a disagreement, and the standard error of the estimate; and how far the probability of winning is from 0.5 in standard units (standard errors).

Standard errors are computed using a normal approximation to the binomial distribution. Generally, much more than 2 standard units means that the difference is statistically significant.

PARTITION

partition [-p q | -m N]

Chooses a random subset of stem.data. Size of subset is specified by either a percentage to retain (-p) or by a subset size (-m).

SELECT-CLASS

select-class -c C [stem]

For examples in stem.data, changes class of examples of anything other than class C to 'default_class', and writes the modified dataset to the standard output.

PPRINT-RULES

pprint-rules [options] stem

Print a RIPPER ruleset to standard output in a human-readable format.

EXAMPLES

Here are some common ways of evaluating RIPPER.

#generate a ruleset

ripper foo

#eliminate redundant conditions

eliminate foo

#test ruleset on foo.test and summarize

predict foo | summarize foo

#save predictions in a file

predict foo > foo-default.pred

#re-run experiment with different RIPPER options

ripper -L2 foo

predict foo > foo-L2.pred

#do a paired test

pair foo-default.pred foo-L2.pred

BUGS

Probably.