Introduction to Weka Experimenter

DATA MINING

REALIZED BY :

Ghassen Daoud

Mehdi Ben Chikha

GL4 - 2 GROUP

Data sets

IRIS :

instances 150

features 4

Weather:

instances 214

features 9

Glass :

instances 14

features 4

Cross validation (folds k = 10)

The data is divided into k subsets.

Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set.

Classification algorithms:

ZeroR: It picks the class value that is the majority in the dataset and gives that for all predictions.

J48: is decision tree algorithm. It is an implementation of the C4.8 algorithm in Java

RandomForest: It builds decision trees on different samples and takes their majority vote for classification.

NaiveBays: are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle,

IBK: k-nearest neighbor's algorithm (k-NN) does not build a model, instead it generates a prediction for a test instance just-in-time. The IBK algorithm uses a distance measure to locate k "close" instances in the training data for each test instance and uses those selected instances to make a prediction. (K=1, K=3, K=5, K=10)

Result analysis:

Accuracy: the accuracy reported is the mean in rackets of those 10 runs (Percent_correct).

Dataset	(1) rules.Ze	(2) trees	(3) trees	(4) bayes	(5) lazy.	(6) lazy.	(7) lazy.	(8) lazy.
iris	(100) 33.33	94.73 v	94.67 v	95.53 v	95.40 v	95.20 v	95.73 v	95.73 v
weather.symbolic	(100) 70.00	47.50	66.00	57.50	61.50	70.50	71.00	70.00
Glass	(100) 35.51	67.58 v	79.72 v	49.45 v	69.95 v	70.02 v	66.04 v	63.26 v
	(v/ /*)	(2/1/0)	(2/1/0)	(2/1/0)	(2/1/0)	(2/1/0)	(2/1/0)	(2/1/0)

Dataset	Best classification algorithm
iris	IBK k=5 & IBK k=10
weather	IBK k=5
Glass	RandomForest

Rank: ranking the algorithms by the number of times a given algorithm beat the other algorithms:

A win, means an accuracy that is better than the accuracy of another algorithm and that the difference was statistically significant.

We can see that RandomForest has the highest win compared to others, it means that RandomForest is potentially contender outperforming out baseline of the zeroR and NaiveBayes.

```
>-<   >   < Resultset
  8   8   0 trees.RandomForest
  3   4   1 lazy.IBk '-K 3 -W 0
  3   4   1 lazy.IBk '-K 1 -W 0
  2   3   1 lazy.IBk '-K 5 -W 0
  2   3   1 trees.J48 '-C 0.25
  0   3   3 lazy.IBk '-K 10 -W
 -4   2   6 bayes.NaiveBayes ''
-14   0  14 rules.ZeroR '' 4805
```

Mean Absolute Error: the MAE reported is the mean in rackets of those 10 runs.

```
Dataset                 (1) rules.Z |  (2) tree  (3) tree  (4) baye  (5) lazy  (6) lazy  (7) lazy  (8) lazy
-------------------------------------------------------------------------------------------------------------
iris                  (100)   0.44 |   0.04 *   0.04 *   0.04 *   0.04 *   0.04 *   0.04 *   0.04 *
weather.symbolic      (100)   0.47 |   0.43     0.41     0.43     0.47     0.43     0.43     0.45
Glass                 (100)   0.21 |   0.10 *   0.10 *   0.15 *   0.09 *   0.10 *   0.11 *   0.12 *
```

Dataset	Worst classification algorithm
iris	ZeroR
weather	ZeroR
Glass	ZeroR

Rank: A win, means an error that is higher than the error of another algorithm and that the difference was statistically significant

We can see that zeroR has the highest win compared to others, it means that ZeroR is the most nonperforming algorithm among the others.

>-<	>	<	Resultset
14	14	0	rules.ZeroR '' 480
4	6	2	bayes.NaiveBayes '
2	5	3	lazy.IBk '-K 10 -W
-1	3	4	lazy.IBk '-K 5 -W
-4	0	4	trees.J48 '-C 0.25
-5	0	5	lazy.IBk '-K 3 -W
-5	0	5	lazy.IBk '-K 1 -W
-5	0	5	trees.RandomForest