Assignment 7

Learning Objectives:

1. Explore a new machine learning algorithm, namely JRIP
2. Practice troubleshooting skills – figuring out what your model is doing

Description:
The data set you will be working with for this assignment was created using LightSIDE. It is a data set where the model is predicting whether a person will be given a job or not based on their resume. The job they are applying for is Wikipedia RFA (which is an administrative position you can read about on Wikipedia if you are interested). All of the basic text features have been extracted for you.

Step-by-Step Guide:

1. Complete the readings through week 10.

2. Use the experimenter to determine whether you get significantly different performance from J48 (tree based learning) and JRIP (rule based learning) when you use a feature selection wrapper that selections the top 50 features on each fold. Give screen shots and explain your results.

   Decision Tees J48
   Time taken to build model: 24.89 seconds

   === Stratified cross-validation ===
   === Summary ===

   Correctly Classified Instances 212 60.9195 %
   Incorrectly Classified Instances 136 39.0805 %
   Kappa statistic 0.0775
   Mean absolute error 0.4054
   Root mean squared error 0.5661
   Relative absolute error 96.0874 %
   Root relative squared error 123.3297 %
   Total Number of Instances 348 === Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.362	0.284	0.355	0.362	0.358	false
0.716	0.638	0.722	0.716	0.719	true

   === Confusion Matrix ===

```
 a   b   <-- classified as
38  67 |   a = false
69 174 |   b = true
```

JRIP

Time taken to build model: 75.59 seconds

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         237               68.1034 %
Incorrectly Classified Instances       111               31.8966 %
Kappa statistic                          0.0876
Mean absolute error                      0.3991
Root mean squared error                  0.469
Relative absolute error                 94.604  %
Root relative squared error            102.17   %
Total Number of Instances              348 === Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
0.171     0.099     0.429     0.171     0.245     false
0.901     0.829     0.716     0.901     0.798     true

=== Confusion Matrix ===

 a   b   <-- classified as
18  87 |   a = false
24 219 |   b = true
```

To determine whether they are statistically significant or not we use Experimenter.

Analysing: Percent_correct
Datasets: 1
Resultsets: 2
Confidence: 0.05 (two tailed)
Date: 11/6/07 11:47 AM

Dataset   (1) meta.Att | (2) meta.

```
                    -------------------------
training_set            (100)   62.86 |   66.61
                    -------------------------
                        (v/ /*) |   (0/1/0)
```

Skipped:

Key:

(1) meta.AttributeSelectedClassifier '-E_\"ChiSquaredAttributeEval_\"_-S_\"Ranker_-T_-1.7976931348623157E308_-N_50\"_-W_trees.J48_--_-C_0.25_-M_2' -5.9518054534879478E18
(2) meta.AttributeSelectedClassifier '-E_\"ChiSquaredAttributeEval_\"_-S_\"Ranker_-T_-1.7976931348623157E308_-N_50\"_-W_rules.JRip_--_-F_3_-N_2.0_-O_2_-S_1' -5.9518054534879478E18

From the experimenter it is clear that J48 and JRip aren't statistically significant different.

3. Troubleshoot your results so that you understand why performance was or was not different between tree and rule based learning. The point here is not to identify where the feature space is weak but to investigate why the algorithms did or didn't perform differently. This can be thought of as a more advanced version of what you did with the Titanic dataset earlier in the semester. It should be doable since you will only be considering around 50 features. You may need to be creative. Now explain your results using what you understand about tree and rule based learning and what you found in your error analysis.

   If there was a significant difference in favor of the rule based approach, I would say it's because of the advantage with respect to representing disjunctions. But I don't really see evidence of an important disjunction in the rule based model that was learned, although that model is far simpler than the tree based model. I don't see a lot of redundancy in the tree based model either. The trend is for trees to be worse, and the tree based model is more complex, so I'd say the tree based model may have slightly overfit the data.

   From the confusion matrix it is obvious that the J48 errors are more spread over, ie "a vs b" confusion is same as "b vs a" confusion. However in the case of JRip, the confusion of "a vs b" is 4 times more than "b vs a". It indicates that JRip is heavily biased towards the class b (true) thus predicting many of the instances which belong to class a (false) as also true.

It is understandable from the JRip class of algorithms which are rule based, and tend to favor majority class more than minority class. In fact that is the case in the training data. The number of instances associated with class true is 243, Vs number of instances associated with class false is 105. The training data given had a skewed distribution. In the case of non-skewed distribution, JRip may perform significantly worse.

4. Are the results surprising given the discussion about tree and rule based learning in the book? Why or why not?

   Surprising: JRip combines many attributes and build up a new feature space using AND, OR operations and hence has the ability to leverage interactions between attributes in a more compact way than J48. Hence, intuitively you would expect JRip to perform better than J48. But this was not the case. Also given that the training data had skewed distribution made it more favorable for JRip like algorithms to work better (because of the default category). A non-significant difference would be expected if there is neither an important disjunction and there are some strong predictors so that the tree based model won't overfit too much.

Deliverables: Write up of your exploration process that includes your write-up from 2-4 above.

Submission Mode:

Submit the assignment to blackboard.