Data Mining and Machine Learning: Coursework 3

# Procedure

First of all, I go to website <http://www.randomizer.org/form.htm> and get my random numbers. Once I got all 8 numbers (4 for training and 4 for test), I use the same set of numbers in entire coursework. Then I write a simple “awk” program (See code below) to run “ers.c”, a program will create 16 commands from 16 combination and run it one at a time and save the results into files. While running the 3rd step, I modify a training and testing file by cutting first 900 instance from testing file and put them in training file then run “awk” program with a new modified training and testing data sets to complete 4th step. Again in 5th step, I moved 1100 more instances from modified testing file from 4th step and put them into modified training file from 4th then run “awk” program again.

|  |
| --- |
| BEGIN {  fpr["1"] = 16;fpr["2"] = 35;fpr["3"] = 49;fpr["4"] = 61;rpc["1"] = 5;rpc["2"] = 21;rpc["3"] = 46;rpc["4"] = 49;  c = 1;  for(i=1;i<5;i++) {  for(j=1;j<5;j++) {  cmd = "./ers 65 2100 3120 10 t2000.txt te2000.txt " fpr[i] " " rpc[j] " 10000 > 2T"fpr[i] "-" rpc[j]".txt";  c++;  system(cmd);  print cmd "FINISHED";  }  }  }  {}  END {} |

# Results

## Step 3

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fields per rules. | Rules per class. | | | | | Fields per rules. | Rules per class. | | | | |
| Train | 5 | 21 | 46 | 49 | Test | 5 | 21 | 46 | 49 |
| 16 | 75 | 98 | 100 | 100 | 16 | 17.8906 | 27.6172 | 34.9219 | 38.2422 |
| 35 | 47 | 92 | 100 | 100 | 35 | 0.703125 | 3.61328 | 6.03516 | 6.30859 |
| 49 | 48 | 93 | 99 | 99 | 49 | 0.0976562 | 0.46875 | 1.66016 | 1.95312 |
| 61 | 44 | 93 | 99 | 100 | 61 | 0.078125 | 0.21484 | 0.527344 | 0.410156 |

Figure1: Result table of Step 3.

## Step 4

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fields per rules. | Rules per class. | | | | | Fields per rules. | Rules per class. | | | | |
| Train | 5 | 21 | 46 | 49 | Test | 5 | 21 | 46 | 49 |
| 16 | 71.6 | 78.9 | 81.4 | 80.9 | 16 | 50.7109 | 53.3649 | 50.1659 | 50.3791 |
| 35 | 27.1 | 32.1 | 48.4 | 49.1 | 35 | 17.1327 | 8.34123 | 9.2654 | 7.18009 |
| 49 | 7.9 | 21.3 | 41.6 | 41.8 | 49 | 1.18483 | 0.592417 | 2.20379 | 0.900474 |
| 61 | 5.8 | 20.3 | 38.5 | 41.2 | 61 | 0.49763 | 0.189573 | 0.63981 | 0.236967 |

Figure2: Result table of Step 4.

## Step 5

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fields per rules. | Rules per class. | | | | | Fields per rules. | Rules per class. | | | | |
| Train | 5 | 21 | 46 | 49 | Test | 5 | 21 | 46 | 49 |
| 16 | 72.5238 | 73.4286 | 76.0952 | 78.8095 | 16 | 65.1923 | 60.4487 | 62.8526 | 63.5256 |
| 35 | 22.0476 | 25.0952 | 35.7619 | 34.619 | 35 | 17.5962 | 12.2756 | 13.8141 | 11.5385 |
| 49 | 6.52381 | 13.2381 | 22.6667 | 24.3333 | 49 | 3.58974 | 1.82692 | 2.33974 | 2.5641 |
| 61 | 2.90476 | 10.619 | 21 | 22.3333 | 61 | 0.0961538 | 0.288462 | 0.673077 | 0.608974 |

Figure3: Result table of Step 5.

## Variation in performance with number of fields per rule.

Performances from step 3, 4 and 5 were greatly affected by size of field per rule. Both training set and testing set performed poorly each time the number of fields per rule was increased. When a size of training set is small as seen in step 3, the effect on training set performance is small if you increase the size of rule per class (high overfitting). The performances on testing set are greatly dropped each time you increased the size of fields in all experiments. Once you increase the size of training set it clearly show that increasing the size of field per rule have no relationship with overfitting but it greatly downgrade prediction performance.

## Variation in performance with number of rules per class.

Number of rules seem to slightly reduce overfitting in step 3 as you can see in Figure1but it is not significant. Furthermore, performances of training set were increased when the size of rules per class is bigger in step3. After the size of training set was increased in step 4 and 5, the effect of number of rules per class is non to be seen in testing set performance but it still increased performances on training set which mean more overfitting. The overall performance on testing set tend to be close together no matter what is size of the rule per class.

## Variation in performance with size of training set.

It is clear that the size of training set help reduced overfitting in training performances within step 4 and 5. Overall performances in training seem to be worsen each time a training set size was increased. On the other hand, overall performances are better and closer to performances of training set which mean less overfitting. Considered the performance in step 5 on training set and test set with a size of field per rules at 16, the results of training set are more than 70% and testing set are more than 60%. The results are much better than the results from step 3.