Data mining Assignment

Roll 205119070

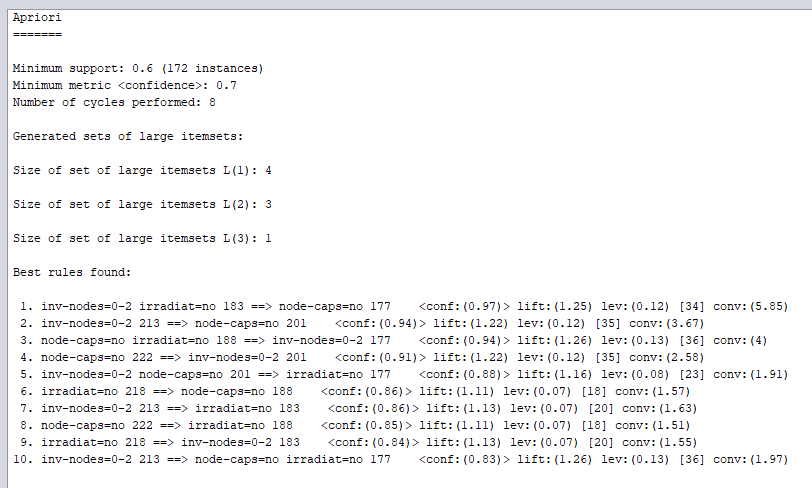
Rahul

Find the association rule for the chosen data set, varying the support from 5 to 20 % in steps of 5 and reducing the confidence from 90 to 70 in steps of 10. Find the number of frequent items and number of association rules generated in a table form as follows:

Sol-Dataset Used breast-cancer.arff

|  |  |  |
| --- | --- | --- |
| Support % | Confidence % | No of frequent item set |
| 5 | 70 | 8 |
| 5 | 80 | 8 |
| 5 | 90 | 17 |
| 10 | 70 | 8 |
| 10 | 80 | 8 |
| 10 | 90 | 17 |
| 15 | 70 | 8 |
| 15 | 80 | 8 |
| 15 | 90 | 17 |
| 20 | 70 | 8 |
| 20 | 80 | 8 |
| 20 | 90 | 17 |

Image for support 20% and confidence 90%



2

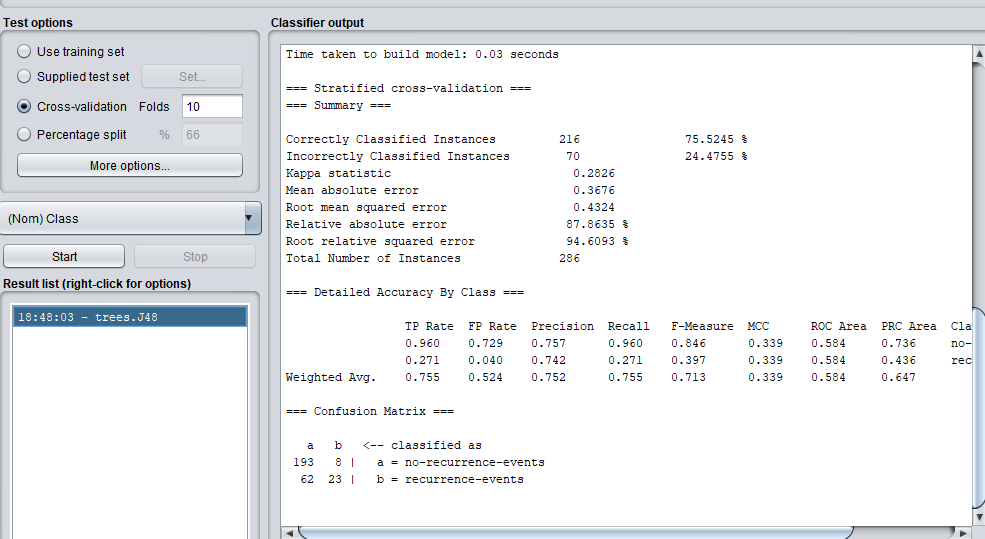
Select the **Id3**classifier. Choose the “Cross-validation” (10 folds) test mode. Run the **Id3**classifier and show the results obtained.

1. How many instances are incorrectly classified?
2. What is the MAE made by the classifier?

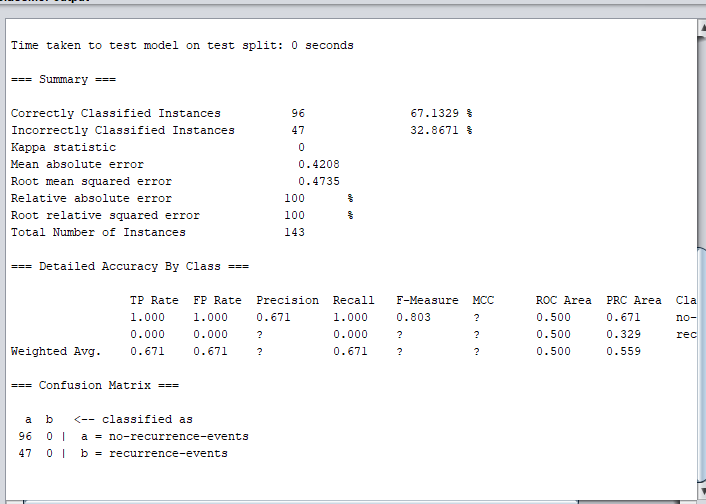
Sol-Dataset Used breast-cancer.arff

a) Incorrectly classified instances-2

b) Mean Absolute Error-0.1429



1. Choose the “Percentage split” (Varying from 50 to 80 in steps of 10% for training and balance for testing) test mode. Run the **Id3**classifier and observe the results shown in the “Classifier output” window.
2. How many instances are incorrectly classified?
3. What is the MAE made by the classifier?
4. Find the confusion matrix for each of the test cases



|  |  |  |  |
| --- | --- | --- | --- |
| Split % | Incorrect classified instances | Mean absolute error | Confusion matrix |
| 50 % | 47 | 0.4208 |  |
| 60% | 39 | 0.4274 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 70% | 32 | 0.4405 |  |
| 80% | 21 | 0.4425 |  |

4. Find the cluster using Euclidean Distance and Manhattan Distance for the chosen dataset using K-Means clustering Algorithm varying K from 3 to 5 and find the number of elements in each cluster.

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=3 | K=4 | K=5 |
| Euclidean distance |  |  |  |
| manhattan |  |  |  |

