Empirical Software Engineering (SE-404)

LAB A1-G2

Laboratory Manual

Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY(DTU)

Shahbad Daulatpur, Bawana Road, Delhi-110042

Submitted to: -

Mr. Sanjay Patidar

Submitted by:-

Name: ASHISH KUMAR

Roll number: 2K18/SE/041

INDEX

Empirical Software Engineering LAB – A1 G2
EXPERIMENT 4

- ASHISH KUMAR
- 2K18/SE/041

Experiment Objective:- Consider defect dataset and perform following feature reduction techniques using WEKA tool. Validate the dataset using 10-cross validation.

a. Correlation based feature evaluation
b. Relief Attribute feature evaluation
c. Information gain feature evaluation
d. Principle Component

Introduction:-

DATASET USED:

- KC2 Dataset in .arff format.
- Author: Mike Chapman, NASA
- Link to dataset: https://datahub.io/machine-learning/kc2#readme

One of the NASA Metrics Data Program defect data sets. Data from software for science data processing. Data comes from McCabe and Halstead features extractors of source code. These features were defined in the 70s in an attempt to objectively characterize code features that are associated with software quality.

Attribute Information

1. loc: numeric McCabe's line count of code
2. v(g) : numeric McCabe "cyclomatic complexity"
3. ev(g) : numeric McCabe "essential complexity"
4. iv(g) : numeric McCabe "design complexity"
5. n: numeric Halstead total operators + operands
6. v: numeric Halstead "volume"
7. l: numeric Halstead "program length"
8. d: numeric Halstead "difficulty"
9. i: numeric Halstead "intelligence"
10. e: numeric Halstead "effort"
11. b: numeric Halstead
12. t: numeric Halstead's time estimator
13. lOCode : numeric % Halstead's line count
14. lOComment : numeric % Halstead's count of lines of comments
15. lOBlank : numeric % Halstead's count of blank lines
16. lOCodeAndComment: numeric
17. uniq_Op : numeric unique operators

18. uniq_Opnd : numeric % unique operands
19. total_Op : numeric % total operators
20. total_Opnd : numeric % total operands
21. branchCount : numeric % Branch Count of the flow graph
22. problems: {false,true} % module has/has not one or more reported defects

Procedure:

1. Open the Weka GUI Chooser.
2. Click the “Explorer” button to launch the Explorer.
3. Open the dataset.
4. Click the “Select attributes” tab to access the feature selection methods

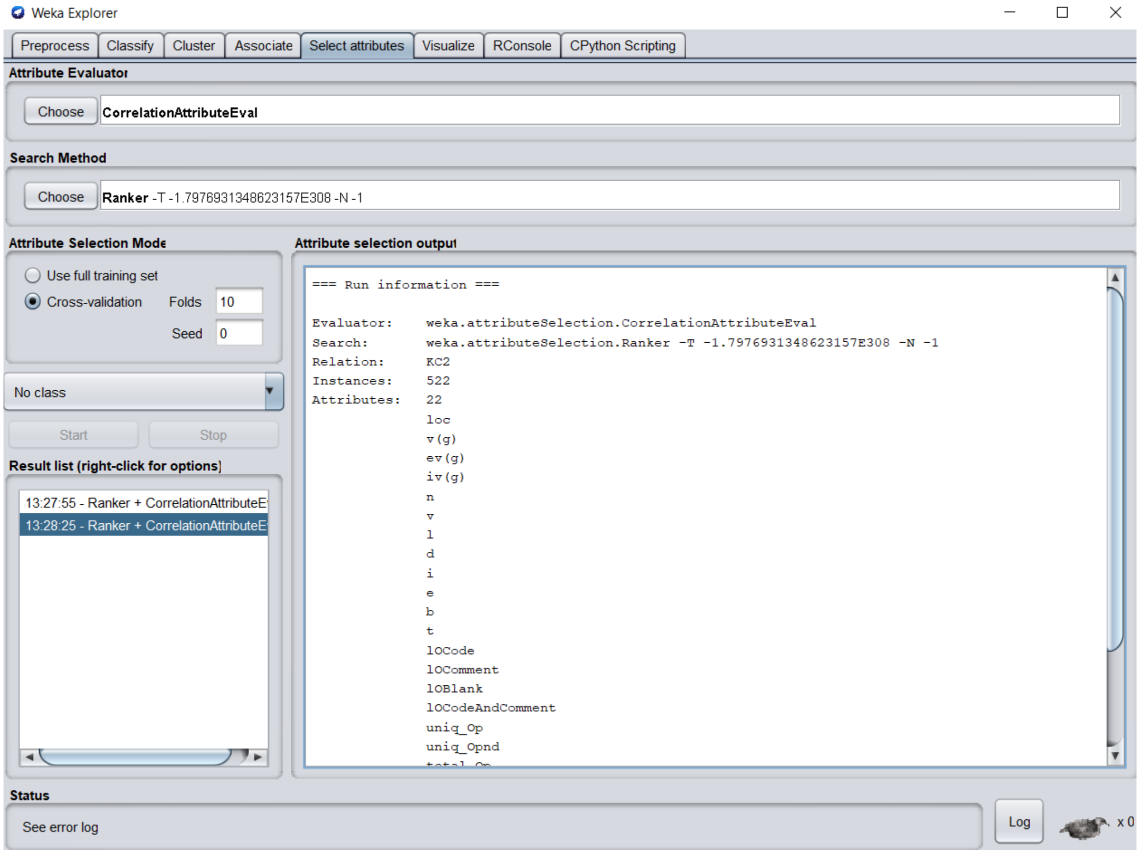
Feature selection is divided into two parts:

- Attribute Evaluator
- Search Method.

Result:-

a) Correlation Based Feature Selection:

A popular technique for selecting the most relevant attributes in your dataset is to use correlation. Correlation is more formally referred to as Pearson’s correlation coefficient in statistics. Weka supports correlation based feature selection with the CorrelationAttributeEval technique that requires use of a Ranker search method.



Attribute Evaluator: CorrelationAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

○ Use full training set
● Cross-validation Folds 10 Seed 0

No class

Result list (right-click for options)

13:27:55 - Ranker + CorrelationAttributeE
13:28:25 - Ranker + CorrelationAttributeE

Attribute selection output

```
=== Run information ===

Evaluator:    weka.attributeSelection.CorrelationAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     KC2
Instances:    522
Attributes:   22
              loc
              v(g)
              ev(g)
              iv(g)
              n
              v
              l
              d
              i
              e
              b
              t
              lOCode
              lOComment
              lOBlank
              lOCodeAndComment
              uniq_Op
              uniq_Opnd
              total_Op
```

Status

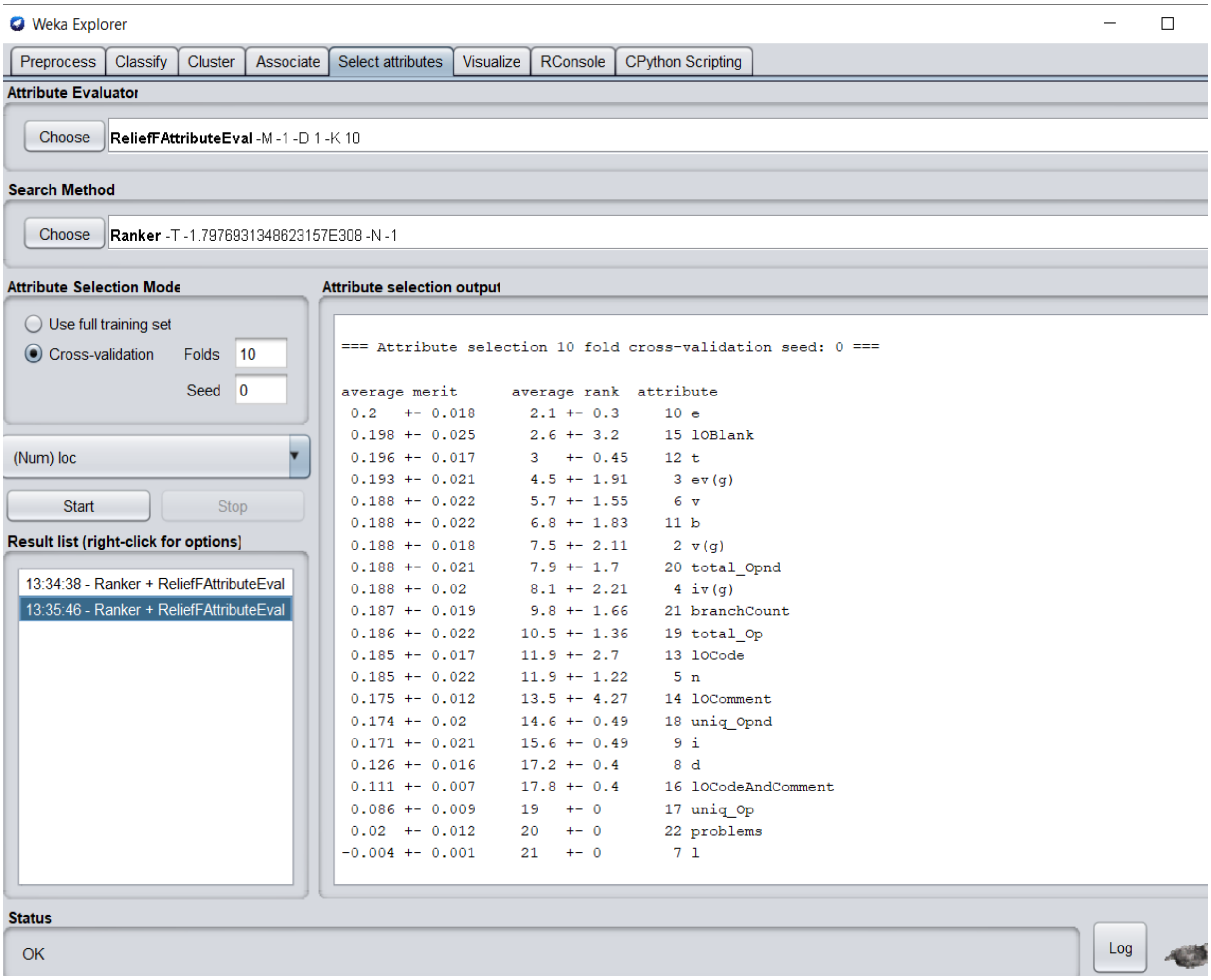
See error log

b) Relief Attribute feature evaluation:

Weka supports correlation based feature selection with ReliefFAttributeEval, the technique that requires use of a Ranker search method.

ReliefFAttributeEval :

Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.



Attribute Evaluator

ReliefFAttributeEval -M -1 -D 1 -K 10

Search Method

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

○ Use full training set

● Cross-validation Folds 10

Seed 0

(Num) loc

Result list (right-click for options)

13:34:38 - Ranker + ReliefFAttributeEval

13:35:46 - Ranker + ReliefFAttributeEval

Attribute selection output

```
=== Attribute selection 10 fold cross-validation seed: 0 ===

average merit      average rank  attribute
 0.2   +- 0.018      2.1 +- 0.3     10 e
 0.198 +- 0.025      2.6 +- 3.2     15 lOBlank
 0.196 +- 0.017      3   +- 0.45    12 t
 0.193 +- 0.021      4.5 +- 1.91     3 ev(g)
 0.188 +- 0.022      5.7 +- 1.55     6 v
 0.188 +- 0.022      6.8 +- 1.83    11 b
 0.188 +- 0.018      7.5 +- 2.11     2 v(g)
 0.188 +- 0.021      7.9 +- 1.7     20 total_Opnd
 0.188 +- 0.02       8.1 +- 2.21     4 iv(g)
 0.187 +- 0.019      9.8 +- 1.66    21 branchCount
 0.186 +- 0.022     10.5 +- 1.36    19 total_Op
 0.185 +- 0.017     11.9 +- 2.7     13 lOCode
 0.185 +- 0.022     11.9 +- 1.22     5 n
 0.175 +- 0.012     13.5 +- 4.27    14 lOComment
 0.174 +- 0.02      14.6 +- 0.49    18 uniq_Opnd
 0.171 +- 0.021     15.6 +- 0.49     9 i
 0.126 +- 0.016     17.2 +- 0.4      8 d
 0.111 +- 0.007     17.8 +- 0.4     16 lOCodeAndComment
 0.086 +- 0.009     19   +- 0       17 uniq_Op
 0.02  +- 0.012     20   +- 0       22 problems
-0.004 +- 0.001     21   +- 0        7 l
```

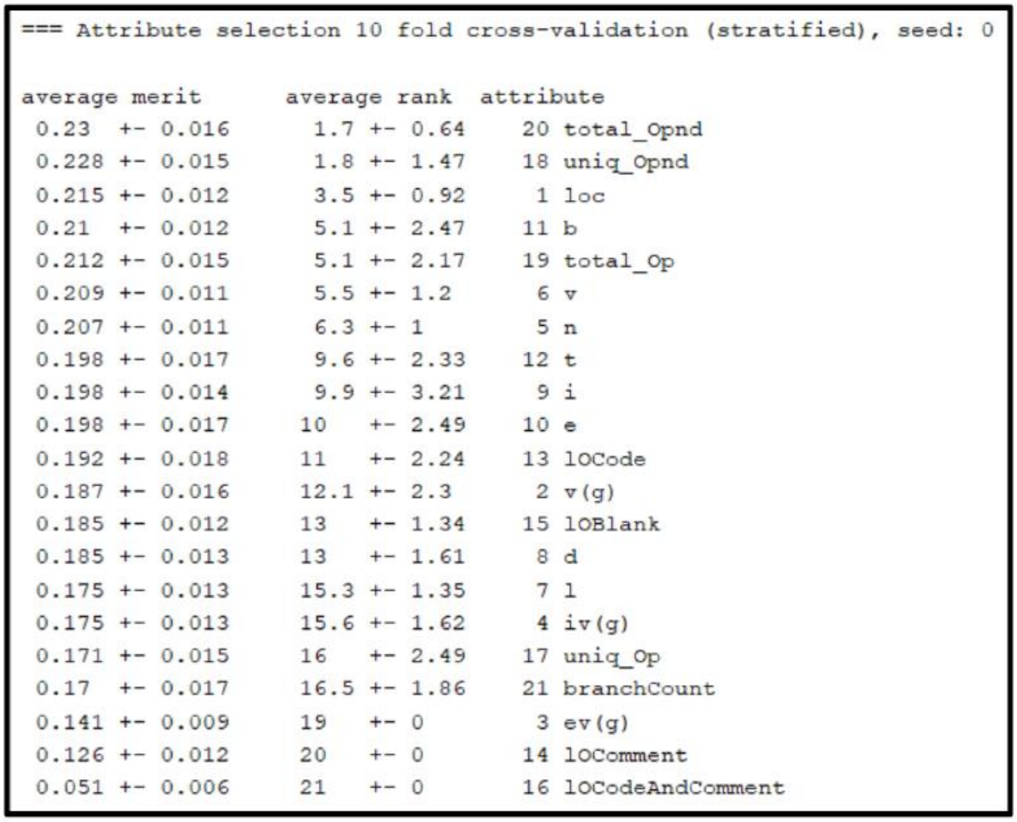
Status

OK

c) Information gain feature evaluation:

Another popular feature selection technique is to calculate the information gain. You can calculate the information gain (also called entropy) for each attribute for the output variable. Entry values vary from 0 (no information) to 1 (maximum information).

Weka supports feature selection via information gain using the InfoGainAttributeEval Attribute Evaluator. Like the correlation technique above, the Ranker Search Method must be used.

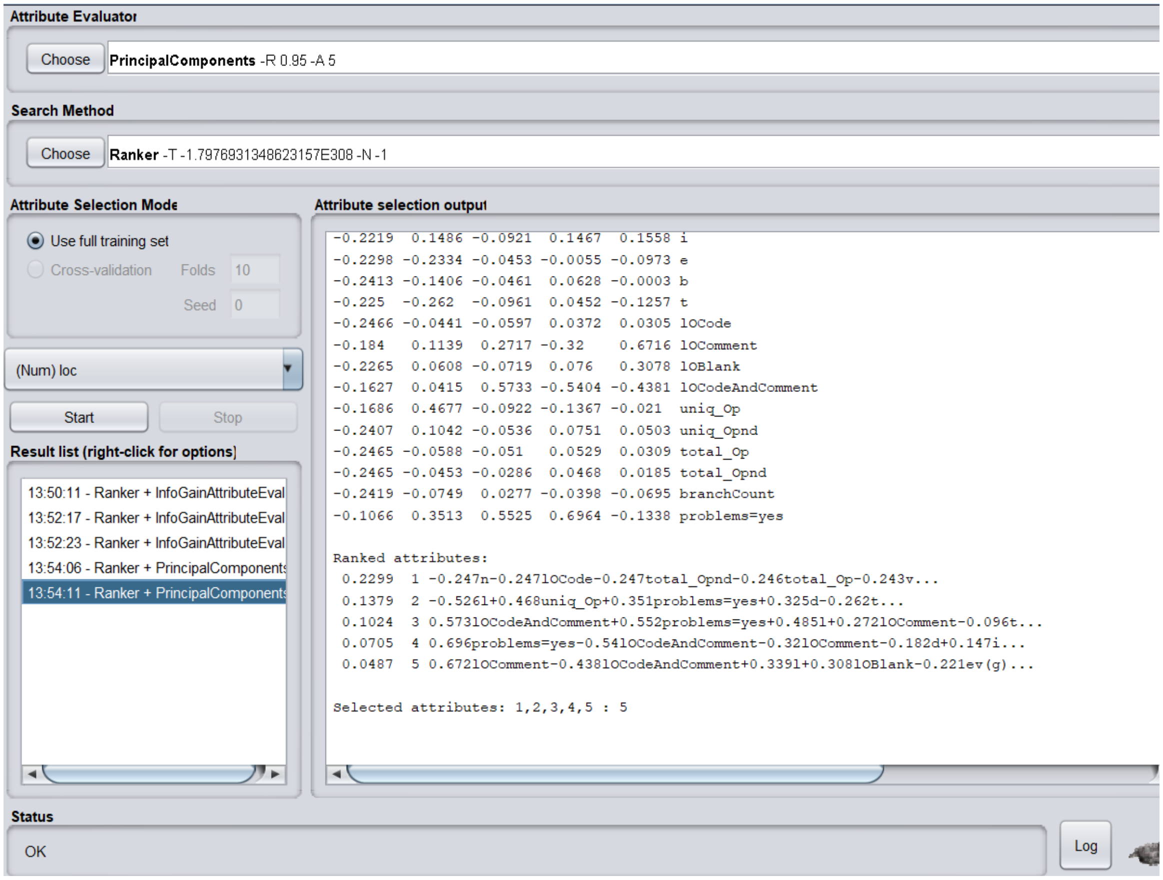


```
=== Attribute selection 10 fold cross-validation (stratified), seed: 0

average merit      average rank  attribute
 0.23  +- 0.016      1.7 +- 0.64    20 total_Opnd
 0.228 +- 0.015      1.8 +- 1.47    18 uniq_Opnd
 0.215 +- 0.012      3.5 +- 0.92     1 loc
 0.21  +- 0.012      5.1 +- 2.47    11 b
 0.212 +- 0.015      5.1 +- 2.17    19 total_Op
 0.209 +- 0.011      5.5 +- 1.2      6 v
 0.207 +- 0.011      6.3 +- 1        5 n
 0.198 +- 0.017      9.6 +- 2.33    12 t
 0.198 +- 0.014      9.9 +- 3.21     9 i
 0.198 +- 0.017     10   +- 2.49    10 e
 0.192 +- 0.018     11   +- 2.24    13 lOCode
 0.187 +- 0.016     12.1 +- 2.3      2 v(g)
 0.185 +- 0.012     13   +- 1.34    15 lOBlank
 0.185 +- 0.013     13   +- 1.61     8 d
 0.175 +- 0.013     15.3 +- 1.35     7 l
 0.175 +- 0.013     15.6 +- 1.62     4 iv(g)
 0.171 +- 0.015     16   +- 2.49    17 uniq_Op
 0.17  +- 0.017     16.5 +- 1.86    21 branchCount
 0.141 +- 0.009     19   +- 0        3 ev(g)
 0.126 +- 0.012     20   +- 0       14 lOComment
 0.051 +- 0.006     21   +- 0       16 lOCodeAndComment
```

d) Principle Component:

Weka Explorer can be used to perform principal components analysis and transformation of the data. It is used in conjunction with a Ranker search.



Attribute Evaluator

PrincipalComponents -R 0.95 -A 5

Search Method

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set
○ Cross-validation Folds 10 Seed 0

(Num) loc

Result list (right-click for options)

13:50:11 - Ranker + InfoGainAttributeEval
13:52:17 - Ranker + InfoGainAttributeEval
13:52:23 - Ranker + InfoGainAttributeEval
13:54:06 - Ranker + PrincipalComponents
13:54:11 - Ranker + PrincipalComponents

Attribute selection output

```
-0.2219  0.1486 -0.0921  0.1467  0.1558 i
-0.2298 -0.2334 -0.0453 -0.0055 -0.0973 e
-0.2413 -0.1406 -0.0461  0.0628 -0.0003 b
-0.225  -0.262  -0.0961  0.0452 -0.1257 t
-0.2466 -0.0441 -0.0597  0.0372  0.0305 lOCode
-0.184   0.1139  0.2717 -0.32    0.6716 lOComment
-0.2265  0.0608 -0.0719  0.076   0.3078 lOBlank
-0.1627  0.0415  0.5733 -0.5404 -0.4381 lOCodeAndComment
-0.1686  0.4677 -0.0922 -0.1367 -0.021  uniq_Op
-0.2407  0.1042 -0.0536  0.0751  0.0503 uniq_Opnd
-0.2465 -0.0588 -0.051   0.0529  0.0309 total_Op
-0.2465 -0.0453 -0.0286  0.0468  0.0185 total_Opnd
-0.2419 -0.0749  0.0277 -0.0398 -0.0695 branchCount
-0.1066  0.3513  0.5525  0.6964 -0.1338 problems=yes

Ranked attributes:
 0.2299  1 -0.247n-0.247lOCode-0.247total_Opnd-0.246total_Op-0.243v...
 0.1379  2 -0.526l+0.468uniq_Op+0.351problems=yes+0.325d-0.262t...
 0.1024  3 0.573lOCodeAndComment+0.552problems=yes+0.485l+0.272lOComment-0.096t...
 0.0705  4 0.696problems=yes-0.54lOCodeAndComment-0.32lOComment-0.182d+0.147i...
 0.0487  5 0.672lOComment-0.438lOCodeAndComment+0.339l+0.308lOBlank-0.221ev(g)...

Selected attributes: 1,2,3,4,5 : 5
```

Status

OK

Learning from experiment:- We have successfully learned about WEKA and Correlation based feature evaluation, Relief Attribute feature evaluation, Information gain feature evaluation and Principle Component. We successfully have used the select attributes feature of WEKA for validating a dataset using 10-cross validation.