EXPERIMENT: 4

(2K17/SE/79 PARV GUPTA)

AIM: 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

THEORY:

DATASET USED:

- KC2 Dataset in .arff format.
- Author: Mike Chapman, NASA
- 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

Feature Selection in Weka:

A central problem in machine learning is identifying a representative set of features from which to construct a classification model for a particular task.
Many feature selection techniques are supported in Weka.

STEPS:

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.

Each section has multiple techniques from which to choose.The attribute evaluator is the technique by which each attribute in the dataset is evaluated in the context of the output variable. The search method is the technique by which to try or navigate different combinations of attributes in the dataset in order to arrive on a shortlist of chosen features. Some Attribute Evaluator techniques require the use of specific Search Methods. For ex, the CorrelationAttributeEval technique used in the next section can only be used with a Ranker Search Method, that evaluates each attribute and lists the results in a rank order.

Both the Attribute Evaluator and Search Method techniques can be configured. Once chosen, click on the name of the technique to get access to its configuration details.

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.You can calculate the correlation between each attribute and the output variable and select only those attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1) and drop those attributes with a low correlation (value close to zero).

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

RESULTS:

```
Evaluator :    weka.attributeSelection.CorrelationAttributeEval
Search    :    weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:      KC2
Instances:     522
Attributes:    22
Evaluation mode:    10-fold cross-validation
```

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

average merit      average rank  attribute
 0.496 +- 0.015      1.4 +- 0.66    17 uniq_Op
 0.49  +- 0.016      2   +- 0.77     8 d
 0.481 +- 0.03       3   +- 0.77    18 uniq_Opnd
 0.473 +- 0.024      3.6 +- 0.66     9 i
 0.411 +- 0.032      5.5 +- 0.5      1 loc
 0.407 +- 0.016      6.3 +- 2.61    15 lOBlank
 0.404 +- 0.034      7   +- 0.45    20 total_Opnd
 0.39  +- 0.037      8.4 +- 1.02    19 total_Op
 0.39  +- 0.037      8.7 +- 0.46    13 lOCode
 0.386 +- 0.035     10   +- 0.45     5 n
 0.377 +- 0.028     11   +- 0.45     2 v(g)
 0.367 +- 0.028     12.5 +- 0.92    21 branchCount
 0.351 +- 0.019     13.1 +- 2.39    14 lOComment
 0.358 +- 0.035     13.2 +- 1.17     4 iv(g)
 0.339 +- 0.037     14.8 +- 0.6     11 b
 0.337 +- 0.04      15.6 +- 1.2      6 v
 0.316 +- 0.012     17.7 +- 0.78     7 l
 0.315 +- 0.038     18.2 +- 0.98     3 ev(g)
 0.306 +- 0.012     18.2 +- 1.17    16 lOCodeAndComment
 0.244 +- 0.023     19.9 +- 0.3     10 e
 0.223 +- 0.029     20.9 +- 0.3     12 t
```

From Classification scores, a ranked list of features is obtained. Experiments with choosing a select number of the highest ranked features and using them with common machine learning algorithms showed that, on average, the top three or more features are as accurate as using the original set. Each Feature's

weight reflects its ability to distinguish among the class values. Features are ranked by weight and those exceed a user-specified threshold are included in the final subset of features. Running this on the KC2 dataset suggests that the attribute (uniq_Op) has the highest correlation with the output class closely followed by the attribute (d).

b) Relief Attribute feature evaluation:

Weka supports correlation based feature selection with ReliefFAttributeEval, the technique that requires use of a Ranker search method. From Classification scores, a ranked list of features is obtained. Each Feature's weight reflects its ability to distinguish among the class values. Features are ranked by weight and those exceed a user-specified threshold are included in the final subset of features.

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.

RESULTS:

Evaluator:	weka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10
Search:	weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:	KC2
Instances:	522

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

average merit      average rank  attribute
 0.05  +- 0.005      1   +- 0        7 l
 0.021 +- 0.003      2   +- 0       17 uniq_Op
 0.011 +- 0.004      3.1 +- 0.3      9 i
 0.009 +- 0.001      4   +- 0.45     8 d
 0.008 +- 0.004      5.3 +- 0.64    18 uniq_Opnd
 0.006 +- 0.001      6.8 +- 3.76    16 lOCodeAndComment
 0.004 +- 0.001      8.4 +- 1.74    15 lOBlank
 0.004 +- 0.001      8.5 +- 3.53    14 lOComment
 0.003 +- 0          9.6 +- 1.85    11 b
 0.003 +- 0.001      9.7 +- 1.42     2 v(g)
 0.003 +- 0.001     10.8 +- 1.25    20 total_Opnd
 0.003 +- 0.001     11.1 +- 1.45    21 branchCount
 0.003 +- 0.001     12.9 +- 1.37     1 loc
 0.003 +- 0.001     13.6 +- 1.02    19 total_Op
 0.002 +- 0.001     14.6 +- 1.11     5 n
 0.002 +- 0.001     15.9 +- 2.02    13 lOCode
 0.002 +- 0.001     16.1 +- 0.94     4 iv(g)
 0.002 +- 0.001     17.8 +- 0.6      6 v
 0.001 +- 0.001     19.3 +- 1        3 ev(g)
 0.001 +- 0         19.6 +- 0.49    10 e
 0.001 +- 0         20.9 +- 0.3     12 t
```

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). Those attributes that contribute more information will have a higher information gain value and can be selected, whereas those that do not add much information will have a lower score and can be removed.

Weka supports feature selection via information gain using the InfoGainAttributeEval Attribute Evaluator. Like the correlation

technique above, the Ranker Search Method must be used.

RESULTS:

```
Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     KC2
Instances:    522
Attributes:   22
```

```
=== 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.

RESULTS:

```
Evaluator:    weka.attributeSelection.PrincipalComponents -R 0.95 -A 5
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     KC2
Instances:    522
Attributes:   22
Evaluation mode:    evaluate on all training data
Search Method:          Attribute ranking.
Attribute Evaluator (unsupervised):    Principal Components Attribute Transformer
```

```
Ranked attributes:
 0.1912   1 -0.241loc-0.241n-0.241lOCode-0.241total_Op-0.241total_Opnd...
 0.1066   2 0.586l-0.505uniq_Op-0.354d+0.259t+0.229e...
 0.0725   3 -0.782lOCodeAndComment-0.414lOComment-0.323l+0.147i-0.145d...
 0.0502   4 -0.599lOComment+0.439lOCodeAndComment-0.416l-0.318lOBlank+0.228ev(g)...
 0.0334   5 0.58 lOComment-0.443i-0.352l-0.278lOBlank-0.26lOCodeAndComment...

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