Handwritten Digit Recognition using Weka and ML

BISHNU POUDEL

Steps used to solve problem

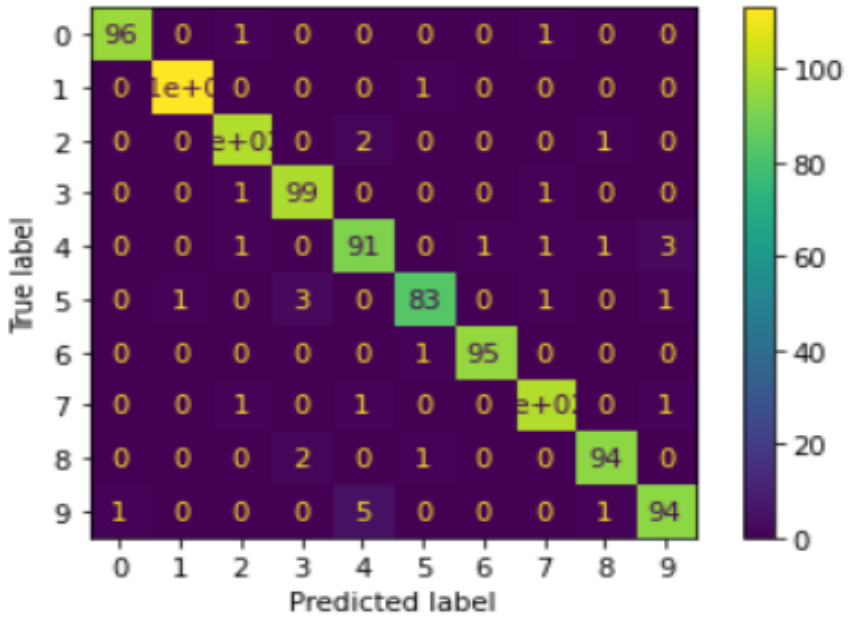
	STEP1	STEP2	STEP3
Tool Used	Solved it using python notebooks in order to find the best algorithm.	Used Weka GUI	Classify in a java code.
Result summary	SVM using rbf kernel and regularization parameter = 1 gave best result. 96% accuracy on test data.	Solved using Image Filters features, then Attribute Selection, then Random Forest Algorithm. 92 % accuracy on test set.	• Converted image to pixels using own code. • Not used to tools like OpenIMAG, Maven. So, still work to do to get used to these. • 99.5 % accuracy on test set.
Further steps	Attempted feature engineering using CV2 library.	SVM gave 91% accuracy on test set.	*Deployment and export of the model remaining.*

Link to python notebook

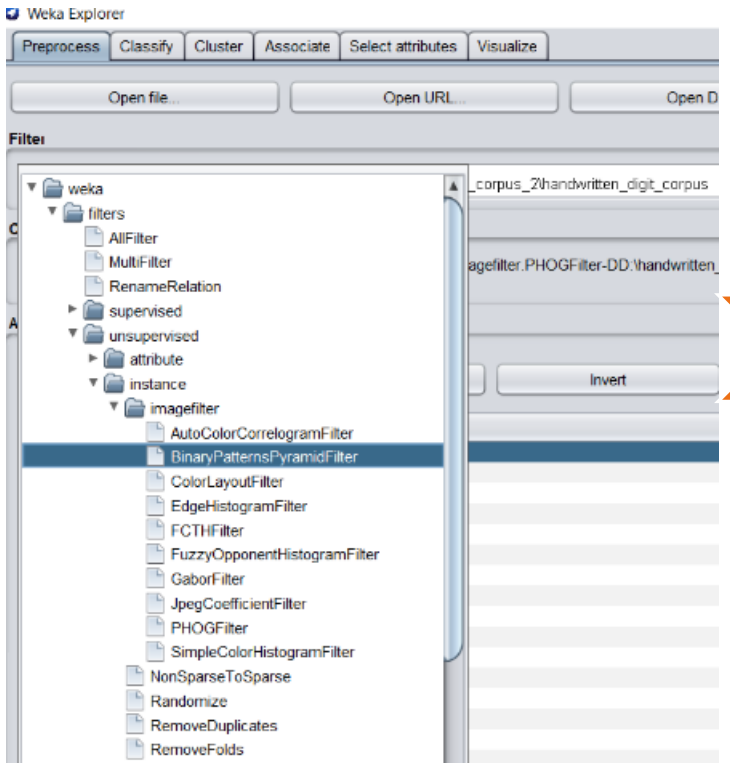
[Link to colab notebook]

```
from sklearn.metrics import plot_confusion_matrix
plot_confusion_matrix(clf, X_test, y_test)
```

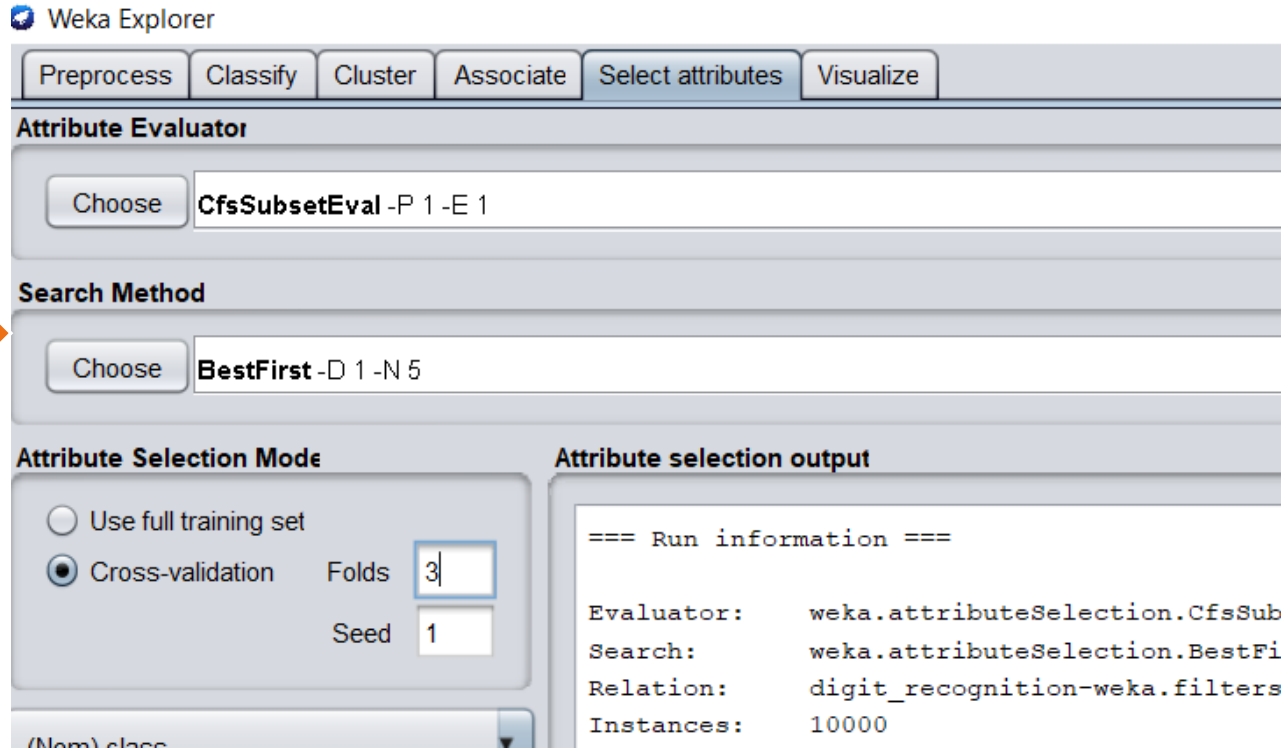
```
<sklearn.metrics._plot.confusion_matrix.ConfusionM
```



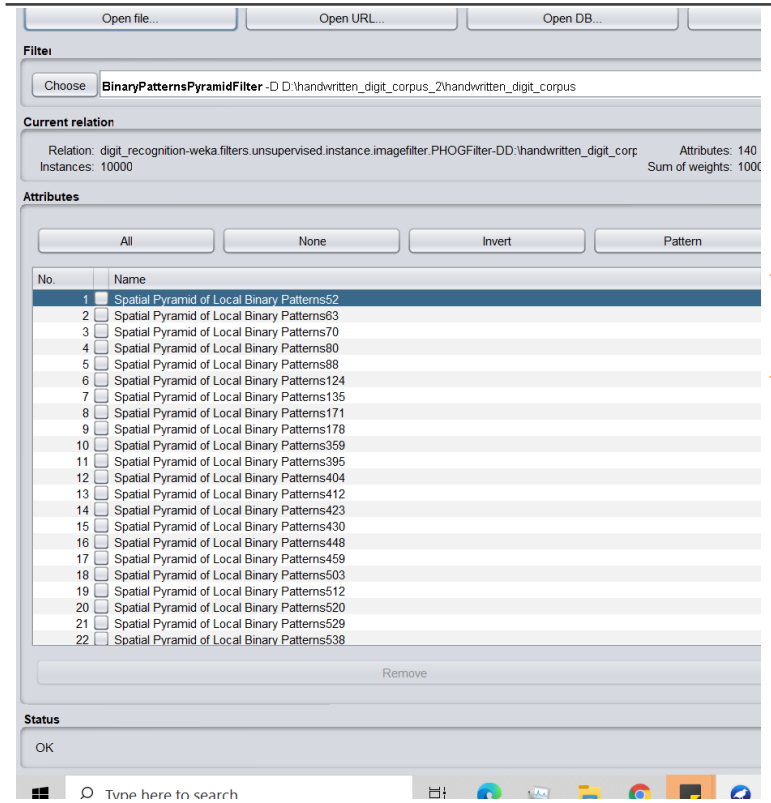
Weka GUI: Generate filter attributes and then Select Attributes



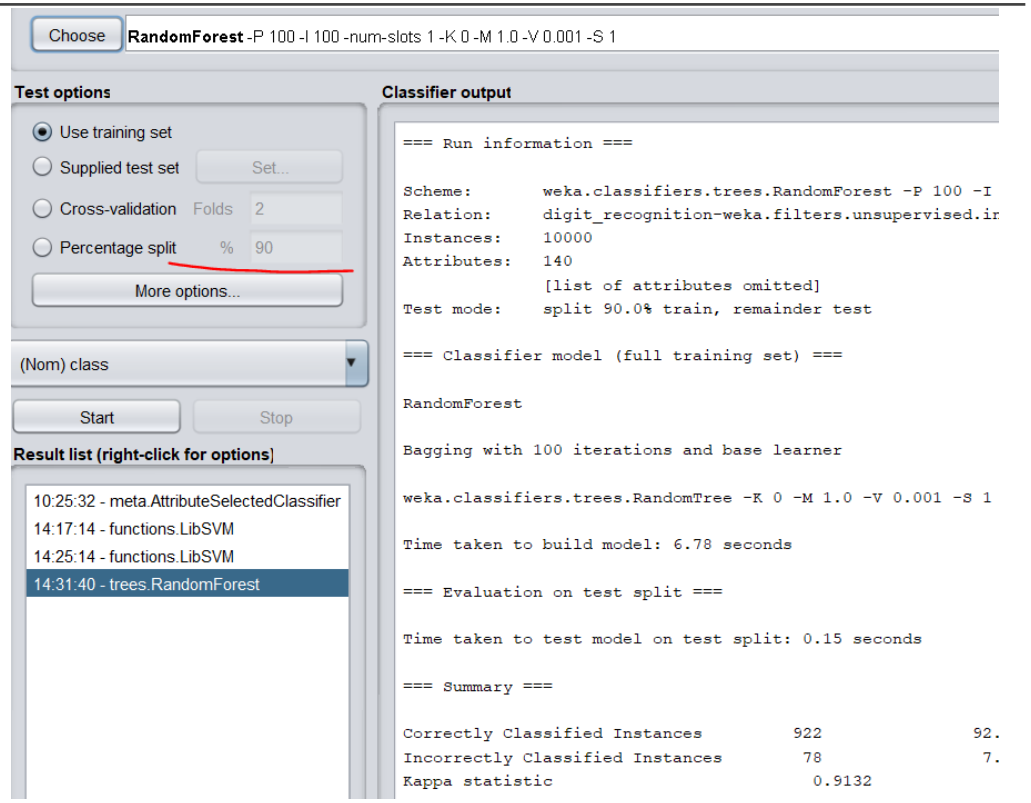
then



Weka GUI: Selected 140 Attributes and Run Random Forest



then



Weka GUI: Random Forest result

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -nu
Relation:     digit_recognition-weka.filters.unsupervised.instance.
Instances:    10000
Attributes:   140
              [list of attributes omitted]
Test mode:    split 90.0% train, remainder test

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not

Time taken to build model: 6.78 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.15 seconds

=== Summary ===

Correctly Classified Instances         922               92.2    %
Incorrectly Classified Instances        78                7.8    %
```

```
=== Confusion Matrix ===

  a   b   c   d   e   f   g   h   i   j   <-- classified as
 87   0   0   0   1   1   0   1   0   0 |   a = 0
  0  99   1   0   2   0   0   1   0   0 |   b = 1
  1   0 106   3   0   0   1   1   2   0 |   c = 2
  2   0   3  82   1   2   0   0   1   2 |   d = 3
  0   0   1   0 103   0   0   1   1   5 |   e = 4
  1   0   1   1   0  96   0   1   2   0 |   f = 5
  1   1   0   0   2   1  88   0   1   0 |   g = 6
  0   1   3   0   3   0   0  89   0   7 |   h = 7
  1   0   0   1   4   3   0   1  74   1 |   i = 8
  1   0   1   0   2   0   0   3   0  98 |   j = 9
```

Weka Java: Link to Code and References

References:

- https://deeplearning.cms.waikato.ac.nz/examples/classifying-mnist/
- https://www.geeksforgeeks.org/image-processing-java-set-2-get-set-pixels/
- https://www.codota.com/code/java/classes/weka.classifiers.evaluation.Evaluation
- https://stackoverflow.com/questions/59345953/weka-how-to-split-dataset-into-70-30-with-stratifiedremovefolds-filter

[Link to GitHub Java code, Weka Jar file and raw skeleton.arff file. Please update the folder paths in the main Class]

```
53          SMO svm = new SMO();
54          svm.buildClassifier(data);
55
56          Evaluation eval = new Evaluation(data);
57           eval.crossValidateModel(svm, data, 3 , new Random(1));
58           //print stats -- do not require to calculate confusion mtx, weka do it!
59           System.out.println(svm);
60           System.out.println(eval.toSummaryString());
61           System.out.println(eval.toMatrixString());
62           System.out.println(eval.toClassDetailsString());
```

Problems Javadoc Declaration Console Terminal Properties

```
terminated> HelloWorld (1) [Java Application] C:\Program Files\Java\jdk-16\bin\javaw.exe  (Mar 22, 2021, 11:46:02 A

Correctly Classified Instances        9090               90.9    %
Incorrectly Classified Instances       910                9.1    %
Kappa statistic                          0.8988
Mean absolute error                      0.1608
Root mean squared error                  0.273
Relative absolute error                 89.3631 %
Root relative squared error             91.017  %
Total Number of Instances            10000

=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j   <-- classified as
 964    0    1    2    0    6    3    1    1    2 |    a = 0
   0 1120    5    2    0    6    0    0    2    0 |    b = 1
   9   17  939   15   14    4    9    9   13    3 |    c = 2
   6    5   27  899    4   32    1   14   15    7 |    d = 3
   2    0   10    2  904    1    9    5    3   46 |    e = 4
   8    9    4   49    6  756   15    5   27   13 |    f = 5
   6    4   23    2    5   19  897    1    1    0 |    g = 6
   6   15   17    6   11    4    1  926    3   39 |    h = 7
   4   17   19   39    9   39    6   13  821    7 |    i = 8
   5    6    5   10   55   11    1   45    7  864 |    j = 9

=== Detailed Accuracy By Class ===
```