# Comparative Analysis of Classification and Regression Algorithms using Weka

Talha Hanif Butt, L18-1864, MS Data Science

National University of Computer and Emerging Sciences, Lahore, Pakistan

***Abstract*—This report presents a comparative analysis of three clustering and three classification algorithms using results produced on three different datasets using Weka.**

## I. Introduction

There has been an increase in the use of machine learning algorithms in the past 5 years and this is because of the performance of certain algorithms like CNNs, RNNs, LSTMs on tasks such as face recognition, speech recognition and autonomous driving. Classification and regression algorithms have been employed on various tasks related to medical, business, education and many other industries and this widespread use allows scientists to easily apply their methods on such problems. This report presents a study of different classification and regression algorithms on different datasets using Weka as a means of learning a new and useful software.

## II. DATASETS

Three datasets have been used in this report for comparing the performance of different algorithms, the details of which are as under:

### *A. Breast Cancer*

The goal in this dataset is to predict recurrence-events given age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad and irradiat. It has 286 instances.

### *B. Airline Delay*

The goal in this dataset is to predict delay given Airline, Flight, To, From, Day, Time and Length. It has 539383 instances.

### *C. Iris Categorization*

The goal in this dataset is to predict the category of iris given Sepal length, Sepal width, Petal length and Petal width. It has 150 instances.

## III. ALGORITHMS

Six algorithms have been used for comparison in this report namely Zero R, Random Forest, Input Mapped Classifier, Farthest First, Filtered Clusterer and Simple K Means.

### *A. Zero R*

Zero R is the simplest classification method which relies on the target and ignores all predictors. Zero R classifier simply predicts the majority category (class). Although there is no predictability power in Zero R, it is useful for determining a baseline performance as a benchmark for other classification methods.

### *B. Random Forest*

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

### *C. Input Mapped Classifier*

Wrapper classifier that addresses incompatible training and test data by building a mapping between the training data that a classifier has been built with and the incoming test instances' structure. Model attributes that are not found in the incoming instances receive missing values, so do incoming nominal attribute values that the classifier has not seen before. A new classifier can be trained or an existing one loaded from a file.

### *D. Farthest First*

The farthest-first traversal of a bounded metric space is a sequence of points in the space, where the first point is selected arbitrarily and each successive point is as far as possible from the set of previously-selected points.

### *E. Filtered Clusterer*

Class for running an arbitrary clusterer on data that has been passed through an arbitrary filter. Like the clusterer, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure.

### *F. Simple K Means*

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the

nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

## IV. Experimental Setup

### *A.* Zero R

| | | |
|---|---|---|
| BatchSize | : | 100 |
| NumDecimalPlaces | : | 2 |

### *B.* Random Forest

| | | |
|---|---|---|
| BagSizePercent | : | 100 |
| BatchSize | : | 100 |
| MaxDepth | : | 0 |
| NumDecimalPlaces | : | 2 |
| NumIterations | : | 100 |

### *C.* Input Mapped Classifier

| | | |
|---|---|---|
| Classifier | : | ZeroR |
| IgnoreCaseForNames | : | True |
| Trim | : | True |

### *D.* Farthest First

NumClusters : 2

### E. Filtered Clusterer

Clusterer : Simple K Means

### *F.* Simple K Means

| | | |
|---|---|---|
| CanopyMaxNumCanopiesToHoldInMemory | : | 100 |
| CanopyMinimumCanopyDensity | : | 2.0 |
| CanopyPeriodicPruningRate | : | 10000 |
| CanopyT1 | : | -1.25 |
| CanopyT2 | : | -1.0 |
| DistanceFunction | : | L1 |
| MaxIterations | : | 500 |
| NumClusters | : | 2 |

## V. RESULTS

### *A.* Zero R on Airline Delay

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 299119 | 55.4558 % |
| Incorrectly Classified Instances | : | 240264 | 44.5442 % |
| Kappa statistic | : | 0 | |
| Mean absolute error | : | 0.494 | |
| Root mean squared error | : | 0.497 | |
| Relative absolute error | : | 100 % | |
| Root relative squared error | : | 100 % | |
| Total Number of Instances | : | 539383 | |

### *B.* Zero R on Breast Cancer

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 201 | 70.2797 % |
| Incorrectly Classified Instances | : | 85 | 29.7203 % |
| Kappa statistic | : | 0 | |
| Mean absolute error | : | 0.4184 | |
| Root mean squared error | : | 0.4571 | |
| Relative absolute error | : | 100 % | |
| Root relative squared error | : | 100 % | |
| Total Number of Instances | : | 286 | |

### *C.* Zero R on Iris Categorization

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 50 | 33.3333 % |
| Incorrectly Classified Instances | : | 100 | 66.6667 % |
| Kappa statistic | : | 0 | |
| Mean absolute error | : | 0.4444 | |
| Root mean squared error | : | 0.4714 | |
| Relative absolute error | : | 100 % | |
| Root relative squared error | : | 100 % | |
| Total Number of Instances | : | 150 | |

### *D.* Random Forest on Airline Delay

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 429108 | 79.5553 % |
| Incorrectly Classified Instances | : | 110275 | 20.4447 % |
| Kappa statistic | : | 0.5843 | |
| Mean absolute error | : | 0.2592 | |
| Root mean squared error | : | 0.3568 | |
| Relative absolute error | : | 52.4584 % | |
| Root relative squared error | : | 71.7871 % | |
| Total Number of Instances | : | 539383 | |

### *E.* Random Forest on Breast Cancer

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 201 | 70.2797 % |
| Incorrectly Classified Instances | : | 85 | 29.7203 % |
| Kappa statistic | : | 0 | |
| Mean absolute error | : | 0.4184 | |
| Root mean squared error | : | 0.4571 | |
| Relative absolute error | : | 100 % | |
| Root relative squared error | : | 100 % | |
| Total Number of Instances | : | 286 | |

### *F.* Random Forest on Iris Categorization

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 199 | 69.5804 % |
| Incorrectly Classified Instances | : | 87 | 30.4196 % |
| Kappa statistic | : | 0.1736 | |
| Mean absolute error | : | 0.3727 | |
| Root mean squared error | : | 0.4613 | |
| Relative absolute error | : | 89.0857 % | |
| Root relative squared error | : | 100.9171 % | |
| Total Number of Instances | : | 286 | |

### *G.* Input Mapped Classifier on Airline Delay

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 299119 | 55.4558 % |
| Incorrectly Classified Instances | : | 240264 | 44.5442 % |
| Kappa statistic | : | 0 | |
| Mean absolute error | : | 0.494 | |
| Root mean squared error | : | 0.497 | |
| Relative absolute error | : | 100 % | |
| Root relative squared error | : | 100 % | |
| Total Number of Instances | : | 539383 | |

***H.*** Input Mapped Classifier on Breast Cancer

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 201 | 70.2797 % |
| Incorrectly Classified Instances | : | 85 | 29.7203 % |
| Kappa statistic | : | 0 | |
| Mean absolute error | : | 0.4184 | |
| Root mean squared error | : | 0.4571 | |
| Relative absolute error | : | 100 | % |
| Root relative squared error | : | 100 | % |
| Total Number of Instances | : | 286 | |

***I.*** Input Mapped Classifier on Iris Categorization

| | | | |
|---|---|---|---|
| Correctly Classified Instances | : | 50 | 33.3333 % |
| Incorrectly Classified Instances | : | 100 | 66.6667 % |
| Kappa statistic | : | 0 | |
| Mean absolute error | : | 0.4444 | |
| Root mean squared error | : | 0.4714 | |
| Relative absolute error | : | 100 | % |
| Root relative squared error | : | 100 | % |
| Total Number of Instances | : | 150 | |

***J.*** Farthest First on Airline Delay

Cluster centroids:

Cluster 0
WN 915.0 CMH STL 2 1120.0 90.0 1
Cluster 1
OO 7781.0 EKO SLC 6 320.0 72.0 0

Clustered Instances

0 301947 ( 56%)
1 237436 ( 44%)

***K.*** Farthest First on Breast Cancer

Cluster centroids:

Cluster 0
60-69 ge40 10-14 0-2 no 1 left left_low no no-recurrence-events
Cluster 1
40-49 premeno 20-24 3-5 yes 2 right right_up yes recurrence-events

Clustered Instances

0 219 ( 77%)
1 67 ( 23%)

***L.*** Farthest First on Iris Categorization

Cluster centroids:

Cluster 0
7.7 3.0 6.1 2.3 Iris-virginica
Cluster 1
4.3 3.0 1.1 0.1 Iris-setosa

Clustered Instances

0 84 ( 56%)
1 66 ( 44%)

***M.*** Filtered Clusterer on Airline Delay

Number of iterations: 7
Within cluster sum of squared errors: 1977647.6109384582

Initial starting points (random):

Cluster 0: WN,2068,DAL,HOU,6,780,60,1
Cluster 1: OO,4683,EUG,SLC,3,780,107,0

Missing values globally replaced with mean/mode

Final cluster centroids:

| | | Cluster# | |
|---|---|---|---|
| Attribute | Full Data | 0 | 1 |
| | (539383.0) | (247360.0) | (292023.0) |
| Airline | WN | WN | OO |
| Flight | 2427.9286 | 1748.8759 | 3003.1247 |
| AirportFrom | ATL | ATL | ATL |
| AirportTo | ATL | ATL | ATL |
| DayOfWeek | 4 | 6 | 3 |
| Time | 802.729 | 837.3786 | 773.3788 |
| Length | 132.202 | 139.8814 | 125.6971 |
| Delay | 0 | 1 | 0 |

Clustered Instances

0 247360 ( 46%)
1 292023 ( 54%)

***N.*** Filtered Clusterer on Breast Cancer

Number of iterations: 3
Within cluster sum of squared errors: 1177.0

Initial starting points (random):

Cluster 0: 50-59,premeno,10-14,0-2,no,2,right,left_up,no,no-recurrence-events
Cluster 1: 40-49,premeno,15-19,0
2,yes,3,right,left_up,no,recurrence-events

Missing values globally replaced with mean/mode

Final cluster centroids:

| | | Cluster# | |
|---|---|---|---|
| Attribute | Full Data | 0 | 1 |
| | (286.0) | (225.0) | (61.0) |
| age | 50-59 | 50-59 | 40-49 |
| menopause | premeno | premeno | premeno |
| tumor-size | 30-34 | 25-29 | 30-34 |
| inv-nodes | 0-2 | 0-2 | 0-2 |
| node-caps | no | no | yes |
| deg-malig | 2 | 2 | 3 |
| breast | left | left | left |
| breast-quad | left_low | left_low | left_low |
| irradiat | no | no | no |

Clustered Instances

0 225 ( 79%)
1 61 ( 21%)

***O.*** Filtered Clusterer on Iris Categorization

Number of iterations: 7
Within cluster sum of squared errors: 62.1436882815797

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:

| | | Cluster# | |
|---|---|---|---|
| Attribute | Full Data | 0 | 1 |
| | (150.0) | (100.0) | (50.0) |
| sepallength | 5.8433 | 6.262 | 5.006 |
| sepalwidth | 3.054 | 2.872 | 3.418 |
| petallength | 3.7587 | 4.906 | 1.464 |
| petalwidth | 1.1987 | 1.676 | 0.244 |
| class | Iris-setosa | Iris-versicolor | Iris-setosa |

Clustered Instances

0 100 ( 67%)
1 50 ( 33%)

***P.*** Simple K Means on Airline Delay

Number of iterations: 7
Within cluster sum of squared errors: 1977647.6109384582

Initial starting points (random):

Cluster 0: WN,2068,DAL,HOU,6,780,60,1
Cluster 1: OO,4683,EUG,SLC,3,780,107,0

Missing values globally replaced with mean/mode

Final cluster centroids:

| | | Cluster# | |
|---|---|---|---|
| Attribute | Full Data | 0 | 1 |
| | (539383.0) | (247360.0) | (292023.0) |
| Airline | WN | WN | OO |
| Flight | 2427.9286 | 1748.8759 | 3003.1247 |
| AirportFrom | ATL | ATL | ATL |
| AirportTo | ATL | ATL | ATL |
| DayOfWeek | 4 | 6 | 3 |
| Time | 802.729 | 837.3786 | 773.3788 |
| Length | 132.202 | 139.8814 | 125.6971 |
| Delay | 0 | 1 | 0 |

Clustered Instances

0 247360 ( 46%)
1 292023 ( 54%)

***Q.*** Simple K Means on Breast Cancer

Number of iterations: 3
Within cluster sum of squared errors: 1177.0

Initial starting points (random):

Cluster 0: 50-59,premeno,10-14,0-2,no,2,right,left_up,no,no-recurrence-events
Cluster 1: 40-49,premeno,15-19,0
2,yes,3,right,left_up,no,recurrence-events

Missing values globally replaced with mean/mode

Final cluster centroids:

| | | Cluster# | |
|---|---|---|---|
| Attribute | Full Data | 0 | 1 |
| | (286.0) | (225.0) | (61.0) |
| age | 50-59 | 50-59 | 40-49 |
| menopause | premeno | premeno | premeno |
| tumor-size | 30-34 | 25-29 | 30-34 |
| inv-nodes | 0-2 | 0-2 | 0-2 |
| node-caps | no | no | yes |
| deg-malig | 2 | 2 | 3 |
| breast | left | left | left |
| breast-quad | left_low | left_low | left_low |
| irradiat | no | no | no |

Clustered Instances

0 225 ( 79%)
1 61 ( 21%)

***R.*** Simple K Means on Iris Categorization

Number of iterations: 7
Within cluster sum of squared errors: 62.1436882815797

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:

| | | Cluster# | |
|---|---|---|---|
| Attribute | Full Data | 0 | 1 |
| | (150.0) | (100.0) | (50.0) |
| sepallength | 5.8433 | 6.262 | 5.006 |
| sepalwidth | 3.054 | 2.872 | 3.418 |
| petallength | 3.7587 | 4.906 | 1.464 |
| petalwidth | 1.1987 | 1.676 | 0.244 |

Clustered Instances

0 100 ( 67%)
1 50 ( 33%)

## VI. COMPARATIVE ANALYSIS

TABLE I
Accuracy Comparison for Classification

| | Airline Delay | Breast Cancer | Iris Categorization |
|---|---|---|---|
| Zero R | 55.45 % | 70.27 % | 33.33 % |
| **Random Forest** | **79.55 %** | **70.27 %** | **69.58 %** |
| Input Mapped | 55.45 % | 70.27 % | 33.33 % |

TABLE 2
Clustered Instance Comparison for Clustering

| | Airline Delay | Breast Cancer | Iris Categorization |
|---|---|---|---|
| Farthest First | 56 %<br>44 % | 77 %<br>23 % | 56 %<br>44 % |
| **Filtered Clusterer** | **46 %**<br>**54 %** | **79 %**<br>**21 %** | **67 %**<br>**33 %** |
| **Simple K Means** | **46 %**<br>**54 %** | **79 %**<br>**21 %** | **67 %**<br>**33 %** |

## VII. CONCLUSION

From Table 1, it can be seen that the accuracy for all the three datasets using Random Forest is the best among Zero R and Input Mapped Classifier. From Table 2, it can be seen that Filtered Clusterer and Simple K Means show similar performance while Farthest First differs and that can be because of the arbitrary starting point or because of the farthest first policy for successive point selection.

## References

http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
http://weka.sourceforge.net/doc.dev/weka/clusterers/FilteredClusterer.html
http://weka.sourceforge.net/doc.dev/weka/classifiers/misc/InputMappedClassifier.html
http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm
https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
http://www.cs.waikato.ac.nz/ml/weka/
http://www.cs.waikato.ac.nz/ml/weka/downloading.html
http://archive.ics.uci.edu/ml/datasets.html
https://en.wikipedia.org/wiki/Farthest-first_traversal