ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Random Forest Modeling For Mice Down Syndrome Through Protein Expression: A Supervised Learning Approach.

RS Kamath[1], TD Dongale[2], Pankaj Pawar[3], and RK Kamat[4]*.

[1]Department of Computer Studies, Chhatrapati Shahu Institute of Business Education and Research, University Road, Kolhapur 416004, Maharashtra, India.
[2]School of Nanoscience and Biotechnology, Shivaji University, Kolhapur, Maharashtra, India
[3]Department of Biotechnology, Shivaji University, Kolhapur – 416004, Maharashtra, India.
[4]Department of Electronics, Shivaji University, Kolhapur – 416 004, Maharashtra, India.

ABSTRACT

We report Random Forest (RF) modeling of expression levels of proteins critical to learning in a mouse model of Down syndrome and delivered detectable signals in the nuclear fraction of the cortex. A random forest is a collection of unpruned decision trees which are often used to model very large datasets. This work exhibits performance evaluation for various RF configurations and compare the classification accuracy. The reported investigation depicts optimum random forest architecture achieved by tuning the number of trees and choice of variables for partitioning the dataset. RF model, thus derived entails 50 trees in the forest with 8 partitioning variables. Moreover the performance of the model is evaluated with reference to Out-of-bag (OOB) estimate of error rate.

Keywords: Random Forest, Protein Expression, Down syndrome, Anti-aging, Decision Trees

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Down syndrome (DS) is one of the most common genetic congenital causes of learning deficits [1]. DS, is a genetic perturbation of considerable complexity due to trisomy of the long (q) arm of human chromosome 21 and the consequent increased level of expression of some subset of the genes it encodes [2]. There are barely any pharmacotherapies available for learning deficits in DS. Presently, protein expression modeling is also turning into an incontestably supportive strategy in microbial cell factories as the learning of the three-dimensional structure of a protein would be a precious guide to take care of issues on protein generation. An interdisciplinary research program has recently been started by the authors with the goal of applying soft computational techniques for protein expressions, enzyme assays, phenotyping, metabolomics and engineering, selecting as well as identifying proteins with a desired activity. Protein expression modeling has been reported by number of researchers in the literature. Centeno et al presented an introduction to comparative modeling with special emphasis on the basic concepts, opportunities and challenges of protein structure prediction [3]. Alireza has depicted collection of Neural Networks to solve class imbalance problem of prediction of secondary protein structure [5]. Benuskova et al have revealed a methodology for using computational neurogenetic modeling to bring new original insights into how genes control the dynamics of brain neural networks [4]. Recently Azizi & Abadehave also reported sequential pattern matching for protein structure prediction [6]. Abd El-Rehim et al. have effectively demonstrated artificial neural network with a back propagation algorithm to identify key biomarkers driving the membership of archival tumor samples [7].

In the backdrop of the research endeavors portrayed above, to the best of our knowledge there are no instances in the literature regarding application of optimum soft computing approach such as the random forest model for classifying mice protein expressions. This algorithm builds multiple decision trees, using a concept called bagging. Bagging is the idea of collecting a random sample of observations into a bag. Each bag of observations is then used as the training dataset for building a decision tree. In the present investigation, we demonstrate the modeling of 77 proteins expression levels measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning. The dataset with 1080 samples of protein is selected for training the forest. The reported experiment is simulated in R and Rattle. R is an open source tool for statistical data processing data mining. Rattle is a graphical data mining package offers GUI for R. The results of the modeling are encouraging and show that the derived RF model efficiently classifies protein samples into the given eight classes with very less error.

The rest of paper is structured as follows; after a brief introduction, second and third sections deals with the infusing theory of mice protein expression and Random Forest respectively. The fourth section outlines our computational details of the RF model with results and discussions. The conclusion at the end discusses aptness of the RF for modelling the mice protein expression.

Mice Protein Expression: Theoretical Considerations

The dataset for RF modeling contains a total of 1080 measurements per protein is taken from UCI data repository. It consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of the cortex. The eight classes of mice are described based on features such as genotype, behavior and treatment. Table 1 lists set of mice classes and corresponding number of observations in the dataset. Fig. 1 shows density of eight classes of mice described in the dataset.

Table 1: Mice protein class details

Mice Protein Class	No. of Observations
c-CS-s	135
c-CS-m	150
c-SC-s	135
c-SC-m	150
t-CS-s	105
t-CS-m	135
t-SC-s	135
t-SC-m	135

ISSN: 0975-8585

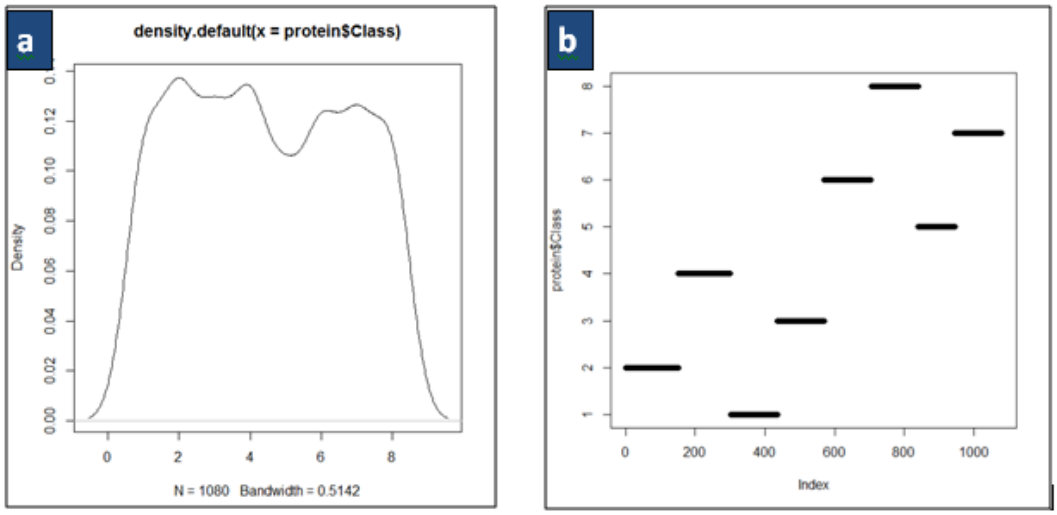

Figure 1: Mice Class Density. Fig (a) Mice class density plot; Fig (b) Mice class plot.

Random Forest: Theoretical considerations

A random forest is collection of unpruned decision trees. It is often used when there is a very large training datasets and a very large number of input variables. This model is typically made up of tens or hundreds of decision trees [10]. These models are generally competitive with nonlinear classifiers.

Random forests is a supervised learning method for classification, that operate by constructing a large number of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees [8, 11]. Here each individual tree will over fit the data and the randomness is in the selection of both observations and choice of variables for partitioning the dataset. The algorithm generates multiple classification and regression trees (CART), and the final classification result is voted among all the trees in the "forest". Random forest classification algorithm is given in fig. 1.

Step 1: Draw n_{tree} bootstrap samples from the original data.

Step 2: For each of the samples, produce an unpruned classification tree, with the following modification:

At each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables.

Step 3: Predict new data by aggregating the predictions of the n_{tree} trees.

Figure 2: Random forest classification algorithm

In the present investigation, the performance of the resulting model is evaluated by OOB [9]. The out-of-bag (OOB) estimate of the error rate is calculated using the observations that are not included in the “bag", the “bag" is the subset of the training dataset used for building the decision tree, hence the “out-of-bag" terminology. This overall measure of accuracy is then followed by a confusion matrix that records the disagreement between the final model's predictions and the actual outcomes of the training observations. Performance of the model can be pictorially represented using Receiver Operating Characteristic (ROC) curve. It plots the true positive rate against the false positive rate. Error plot is useful for deciding optimal number of trees to build. Plot error rate progressively for the number of trees built.

ISSN: 0975-8585

COMPUTATIONAL DETAILS, RESULTS AND DISCUSSIONS

This section explores details of experiment conducted for the classification of mice protein expressions with different RF architectures. R and Rattle were used to analyze model structure, number of trees in the forest and choice of variables for partitioning the dataset [12]. We used the training data set for the parameter adjustment of model whereas validation set to control learning process. We carried out performance evaluation for various RF configurations and compare the classification accuracy. RF builds many decision trees using random subset of data and variables. We used the RANDOMFOREST package for classification by random forest classifiers. For classification, the corresponding method implements Breiman's random-forest algorithm discussed elsewhere in the literature [7]. The said method is proven for assessing proximities among data points in unsupervised mode.

In the present investigation, model is tuned with two parameters n_{tree} and n_{try} to get optimized forest architecture. The parameter n_{tree} specifies how many trees are to be built to populate the random forest where as n_{try} specifies the number of variables that will be considered at any time in deciding how to partition the dataset. We have demonstrated RF modeling per variation in n_{tree} and n_{try}. The entire experiment is summarized in table 2. We have varied value for n_{tree} from 20 to 200 and value for n_{try} from 5 to 15. Table 2 shows performance of RF model with reference to OOB estimate of error rate. Random forest has selected 756 observations randomly to build the model. We have exploited error plot and ROC curve as useful diagnostic tool for our random forest modeling. Figure 3(a-d) presents error plots for the execution of RF models per variation in n_{tree} and n_{try}. The plot reports the accuracy of the forest of trees (in terms of error rate on the y-axis) against the number of trees that have been included in the forest (the x-axis). Figure 4(a-d) presents the ROC curves for different architectures based on the out-of-bag predictions for each observation in the training dataset. The performance of the resulting random forest model tends not to degrade as the number of trees increases, though computationally it takes longer time and implies more inherent complexity to use when scoring, and often there is little to gain from adding too many trees to a forest.

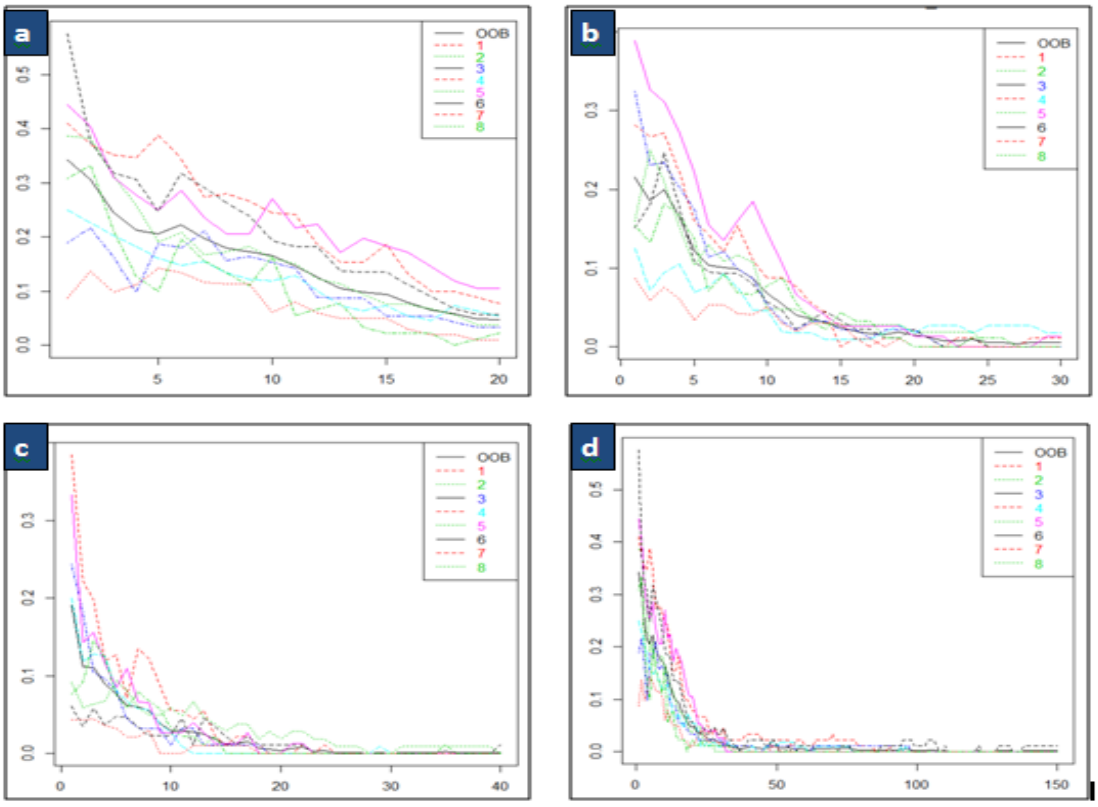

Figure 3: Error plots for RF models per variation in n_{tree} and n_{try}. Fig.(a) represents mean square error of RF model with n_{tree} is 20 and n_{try} is 5; Fig.(b) represents mean square error of RF model with n_{tree} is 30 and n_{try} is 12; Fig.(c) represents mean square error of RF model with n_{tree} is 40 and n_{try} is 15; Fig.(d) represents mean square error of RF model with n_{tree} is 150 and n_{try} is 5.

ISSN: 0975-8585

Table 2: Performance evaluation for accuracy of Random forest Configurations

	OOB estimate of error rate			
No. of Variables → / No. of Tree ↓	5	8	12	15
20	4.76%	1.32%	1.46%	0.4%
25	2.78%	0.53%	0.53%	0.13%
30	1.72%	0.13%	0.53%	0
35	0.66%	0.13%	0.13%	0.26%
40	0.79%	0.4%	0	0.13%
45	0.53%	0.13%	0%	0.26%
50	0.79%	0	0	0.13%
75	0.66%	0	0	0
100	0.13%	0	0	0
150	0.13%	0	0	0
200	0	0	0	0

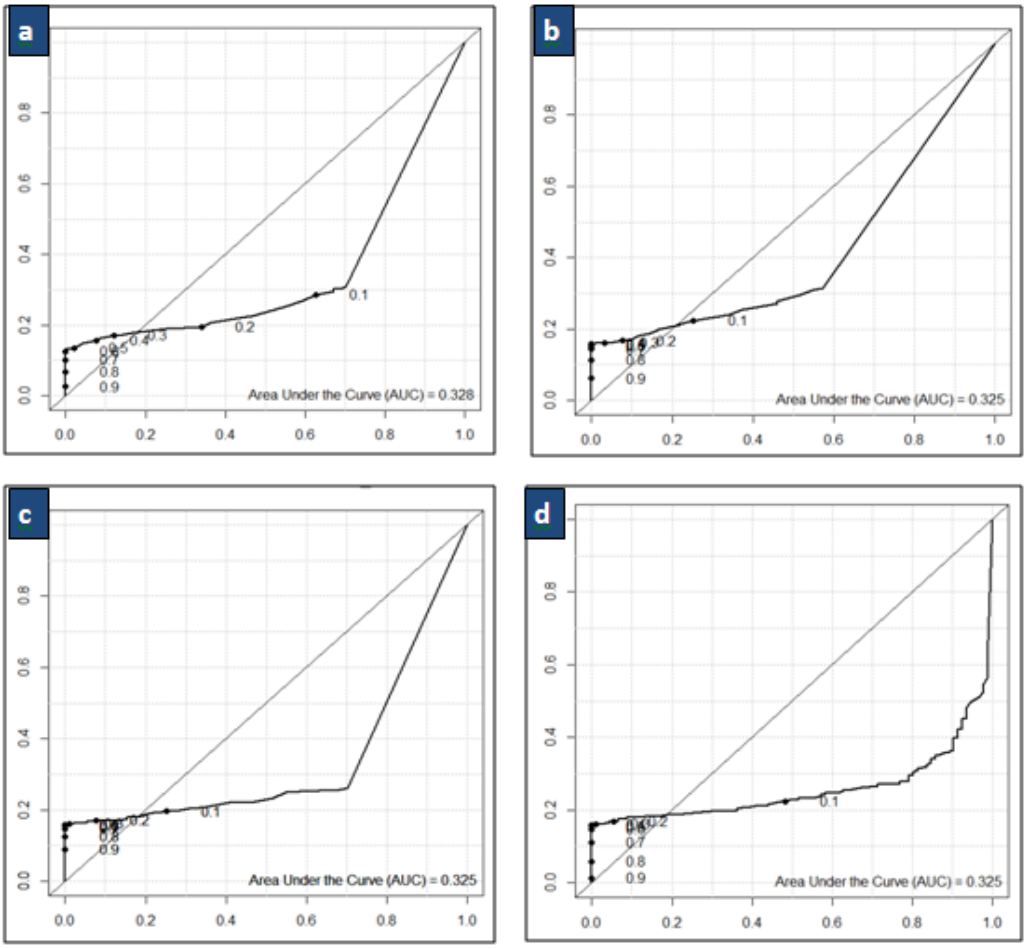

Figure 3: the ROC curves for RF models per variation in n_{tree} and n_{try} based on the out-of-bag predictions for each observation in the training dataset. Fig.(a) represents ROC curve of RF model with n_{tree} is 20 and n_{try} is 5; Fig.(b) represents ROC curve of RF model with n_{tree} is 30 and n_{try} is 12; Fig.(c) represents ROC curve of RF model with n_{tree} is 40 and n_{try} is 15; Fig.(d) represents ROC curve of RF model with n_{tree} is 150 and n_{try} is 5.

The optimized RF architecture selected for the modeling of Mice Protein Expression entails 50 trees in the forest with 8 partitioning variable. RF Model has selected 756 observations randomly to build the forest. Fig. 4 details the structure of optimized RF model selected. The performance of RF modeling pertaining to this is shown in figure 5(a-b). In this case, OOB estimate error rate found to be 0. The dataset with 1080 samples of protein is selected for RF modeling. The selected RF model demonstrates importance of variables and based

ISSN: 0975-8585

on this 756 observations randomly selected for modeling of mice protein expressions. Fig. 5 shows the relative importance of the variables of the dataset taken under the study.

```
Summary of the Random Forest Model
==================================

Number of observations used to build the model: 756
Missing value imputation is active.

Call:
 randomForest(formula = as.factor(Class) ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 50, mtry = 8, importance = TRUE, replace = FALSE, na.action = na.roughfix)

               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 8

        OOB estimate of  error rate: 0%
Confusion matrix:
   1   2   3   4   5   6   7   8 class.error
1 91   0   0   0   0   0   0   0           0
2  0 106   0   0   0   0   0   0           0
3  0   0  92   0   0   0   0   0           0
4  0   0   0 110   0   0   0   0           0
5  0   0   0   0  76   0   0   0           0
6  0   0   0   0   0  89   0   0           0
7  0   0   0   0   0   0 101   0           0
8  0   0   0   0   0   0   0  91           0
```

Figure 4: Textual representation of selected RF model

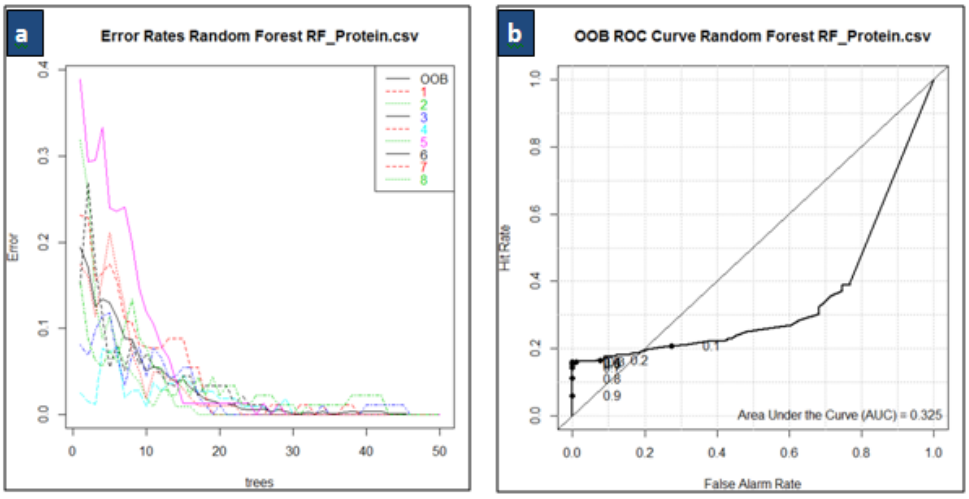

Figure 5: Performance of selected RF model with n_{tree} is 50 and n_{try} is 8; Fig(a) represents mean square error plot; Fig(b) represents ROC curve based OOB

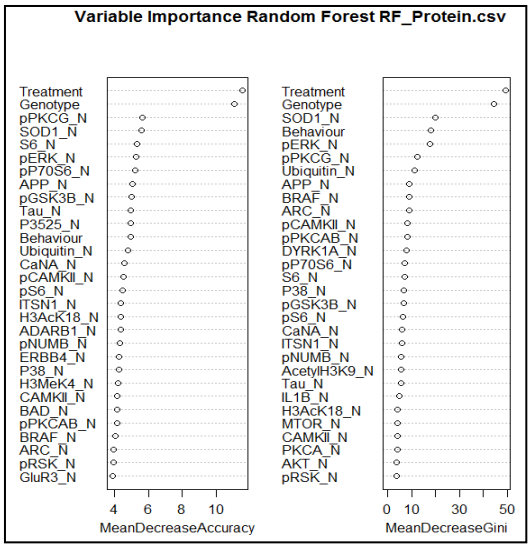

Figure 5: Measure of variable importance

Thus derived RF architecture efficiently classifies new protein samples with very less error. We have tested model with known protein samples. Fig. 6 shows the result obtained in terms of error matrix by applying the test dataset on the derived RF model. Result concludes that RF modeling is a suitable approach since the resulting analysis is much more accurate and precise.

```
Error matrix for the Random Forest model on RF_Protein.csv (counts):

      Predicted
Actual 1 2 3 4 5 6 7 8
     1 5 0 0 0 0 0 0 0
     2 0 5 0 0 0 0 0 0
     3 0 0 5 0 0 0 0 0
     4 0 0 0 5 0 0 0 0
     5 0 0 0 0 5 0 0 0
     6 0 0 0 0 0 5 0 0
     7 0 0 0 0 0 0 5 0
     8 0 0 0 0 0 0 0 5

Error matrix for the Random Forest model on RF_Protein.csv (proportions):

      Predicted
Actual    1    2    3    4    5    6    7    8 Error
     1 0.12 0.00 0.00 0.00 0.00 0.00 0.00 0.00     0
     2 0.00 0.12 0.00 0.00 0.00 0.00 0.00 0.00     0
     3 0.00 0.00 0.12 0.00 0.00 0.00 0.00 0.00     0
     4 0.00 0.00 0.00 0.12 0.00 0.00 0.00 0.00     0
     5 0.00 0.00 0.00 0.00 0.12 0.00 0.00 0.00     0
     6 0.00 0.00 0.00 0.00 0.00 0.12 0.00 0.00     0
     7 0.00 0.00 0.00 0.00 0.00 0.00 0.12 0.00     0
     8 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.12     0
```

Figure 6: Execution result of RF model on test dataset

CONCLUSION

In the present paper, we have reported modeling of expression levels 77 proteins which are considered critical to learning in the mouse model of Down syndrome using Random forest technique. The dataset with 1080 samples of protein were selected for aforesaid modeling. The present investigation demonstrated optimum RF architecture by varying its various attributes such as number of trees and choice of variables for partitioning the dataset. The resulted RF architecture entails 50 trees in the forest with 8 partitioning variable. RF Model has selected 756 observations randomly to build the forest. Thus derived RF model efficiently classifies protein samples into the given eight classes with very less error. The result suggests that the RF has the potential to exhibit as the best tool for modeling of protein samples. Authors are in a process to adopt the technique for prediction, modeling and designing of useful proteins for anti-aging drug design.

ACKNOWLEDGEMENT

We gratefully acknowledge the dataset obtained from the UCI Machine Learning Repository, specifically the Mice Protein Expression Data Set available at http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression

References

[1] Irving C, Basu A, Richmond S, Burn J, Wren C. Eur J Hum Genet 2008; 16(11): 1336–40.
[2] Wiseman F, Alford K, Tybulewicz V, Fisher E. HumMolGenet 2009; 18(R1): 75–83.
[3] Centeno N, Planas-Iglesias J, Oliva B. Microb Cell Fact 2005; *4*(1): 20.
[4] Benuskova L, Jain V, Wysoski S, Kasabov N. Int. J. Neur. Syst. 2006; 16(03): 215-226.
[5] Alirezaee M. Int. J. Artificial Intelligence & Applications 2012; 3(6): 9-20.
[6] Azizi M, Abade M. J. Artificial Intelligence & Applications 2015; 6(4): 31-42.
[7] Abd El-Rehim D, Ball G, Pinder S, Rakha, E, Paish C, Robertson J. et al. International Journal Of Cancer 2005; 116(3): 340-350.
[8] Breiman L. Machine Learning 2001; 45(1): 5-32.
[9] Graham W. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer, UK, 2011, pp. 50-53.
[10] Kamath R, Kamat R. Educational Data Mining with R and Rattle, River Publishers, Netherland, 2016, pp. 30-32
[11] Kamath R, Kamat R. Int. J. Information Technology, Modeling and Computing 2016; 4(1): 19-30.
[12] Andy L, Matthew W. R News 2002; 2(3) 2-4.