ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Random Forest Modelling for Cardiotocography Data: A Case Study on Machine Learning with SparkR.

RS Kamath[1]* and RK Kamat[2]

[1]Department of Computer Studies, ChhatrapatiShahu Institute of Business Education and Research, University Road, Kolhapur 416004
[2]Department of Electronics, Shivaji University, Kolhapur – 416 004

ABSTRACT

We present Random Forest modelling of fetal states by exploring Cardiotocography. Cardiotocography comprises of fetal heart rate and tocographic estimations, is utilized to assess fetal prosperity. Random forest is a sort of ensemble learning technique, where a gathering of frail models consolidate to shape an effective model. Present study shows execution estimation of different arbitrary random forest setups and compares the classification precision. The reported study portrays optimal architecture accomplished by tuning the quantity of trees and decision of variables for dividing the dataset. A classification model, in this way inferred involves 400 trees in the forest with 7 dividing variables. Beside the performance of the model is evaluated with reference to mean square error rate. The entire experiment is carried out in SparkR and RStudio software platform.

Keywords: random forest, CTGs, classification, fetal states, R package, RStudio

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Machine learning is a technique for data analysis that mechanizes scientific model building. Apache Spark is an open source big data processing framework worked around pace, convenience, and sophisticated analytics. SparkR, an R bundle that provide a frontend to Apache Spark and uses Spark's distributed computation engine to empower huge scale data analysis from the R shell [9]. Machine learning plays a vital role in many applications [14]. In the proposed investigation, authors have reported random forest modeling of fetal states by exploring cardiotocography. Cardiotocography (CTG) comprises of fetal heart rate (FHR) and tocographic (TOCO) estimations, is utilized to assess fetal prosperity [12]. It utilizes ultrasound waves to gauge the same. The CTG is demonstrated since 27 weeks of pregnancy and it quantifies heart action, uterine compression and fetal development. FHR patterns are seen by obstetricians amid the procedure of CTG investigation [13]. Consequences of the CTG permit perceiving of three essential distinctive fetal states, for example, normal, suspect and pathological. The obtained data is important to envision awfulness of the fetus and gives an open door for early intercession preceding occurrence of a perpetual disability to the incipient organism.

Literature review reveals that there are a few reported occurrences of utilizing the machine learning approaches as a part of the field of CTG information investigation [16-18]. Kamath and Kamat have reported random forest modeling of displaying of fetal morphologic patterns by investigating CTG information and inferred ideal RF design by changing its different properties [1]. Thus derived RF model effectively orders CTG tests into the given ten morphologic example classes with less error. Thomas et al have reported random forest algorithm for automatic recognition of three fundamental diverse fetal states, for example, normal, suspect and pathological [2]. This framework particularly utilized as a part of pre-birth care as a support decision system. Sahin and Subasi have reported the exploration that assesses the exhibitions of different machine-learning strategies on the CTG information [3]. The exploration uncovered that grouping is important to anticipate infant wellbeing, particularly for the basic cases. Sundar et al have planned artificial neural network model for the classification of CTG data [4]. This classifier was fit for distinguishing Normal, Suspicious and Pathologic condition with fewer errors. However another paper by Karabulut and Ibrikci have uncovered a PC based methodology for breaking down CTG data by utilizing decision tree and different other machine learning calculations [5]. Out of which decision tree adds to an official choice of the framework with precision 95.01%. Magenes et al have portrayed neural classifiers to separate among fetal behavioral states on the premise of CTG signals [6]. These classifiers are fed by files coerced from fetal heart rate signal. Research affirmed promising execution towards the expectation of fetal behavioral states on the arrangement of gathered FHR signals.

Consequently, the global situation of demonstrating portrays the scientists endeavoring hard to turn out with a sweeping model with the end goal of investigation and experimentation of CTG information. In the scenery of the exploration tries depicted over, the present paper reports the random forest based methodology for displaying fetal states through CTG Data. The dataset comprises of estimations of fetal heart rate (FHR) and uterine constriction (UC) highlights on CTG information ordered by master obstetricians. The dataset with 2126 examples of fetal CTGs is chosen for demonstrating [7]. The reported experiment is simulated in RStudo and SparkR environment. Random forest is a versatile machine learning approach assembles multiple decision trees, utilizing an idea called bagging [8]. The consequences of the demonstrating are empowering and demonstrate that the determined RF show proficiently classifies CTG data into the given three classes with less error.

The rest of paper is structured as follows; after a brief introduction, second section manages the materials and strategies explored in the present examination. The third segment traces our computational subtle elements of the RF model with results and discourses. The conclusion toward the end talks about inclination of the RF for displaying the fetal behavioral states.

MATERIALS AND METHODS

The dataset for RF demonstrating contains 2126 specimens of fetal CTGs is taken from UCI information repository [7]. It comprises of estimations of FHR and UC highlights on Cardiotocograms. The CTGs were classified by master obstetricians and characterization was both concerning morphologic patterns and to a fetal state. Present exploration reports examination of CTGs information for classifying it into three classes of

ISSN: 0975-8585

fetal states. Table 1 records set of classes and corresponding number of observations in the dataset. Fig. 1 indicates density of these classes depicted in the dataset.

Table 1: Fetal morphologic patterns class details of CTG data

Abbreviation	NSP – Class Detail	No. of Observations
1	Normal	1655
2	Suspect	295
3	Pathologic	176

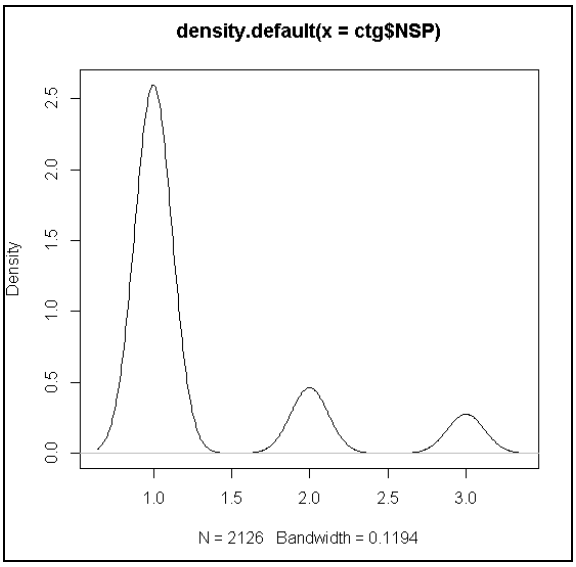

Fig 1: Three classes of fetal state density

The present investigation is carried out in SparkR and RStudio software platform. R is a popular statistical programming language with a number of extensions that support data processing and machine learning tasks [11]. However, interactive data analysis in R is usually limited as the runtime is single-threaded and can only process data sets that fit in a single machine's memory. SparkR is an R package that provides a light-weight frontend to use Apache Spark from R [9].It provides a distributed data frame implementation that supports operations on large datasets. It also supports distributed machine learning using MLib. It exposes the Spark API through the Resilient Distributed Dataset class and allows users to interactively run jobs from the R shell on a cluster. RStudio is a free and open-source integrated development environment (IDE) for R [10].

In the present investigation we have utilized random forest modeling for classifying CTGs data in to three classes of fetal states. The model is imagined as a Multi-Input Single-Output arrangement. It works essentially with 21 inputs viz. estimations of FHR and UC highlights. Fetal state class is considered as a yield variable. It works by generating multiple trees as opposed to a single tree in decision tree model [8]. To classify a new object based on attributes, each tree gives a classification and is called as the tree "votes" for that class and the forest chooses the classification having the most votes. The control flow logic of entire experiment is depicted in figure 2.

ISSN: 0975-8585

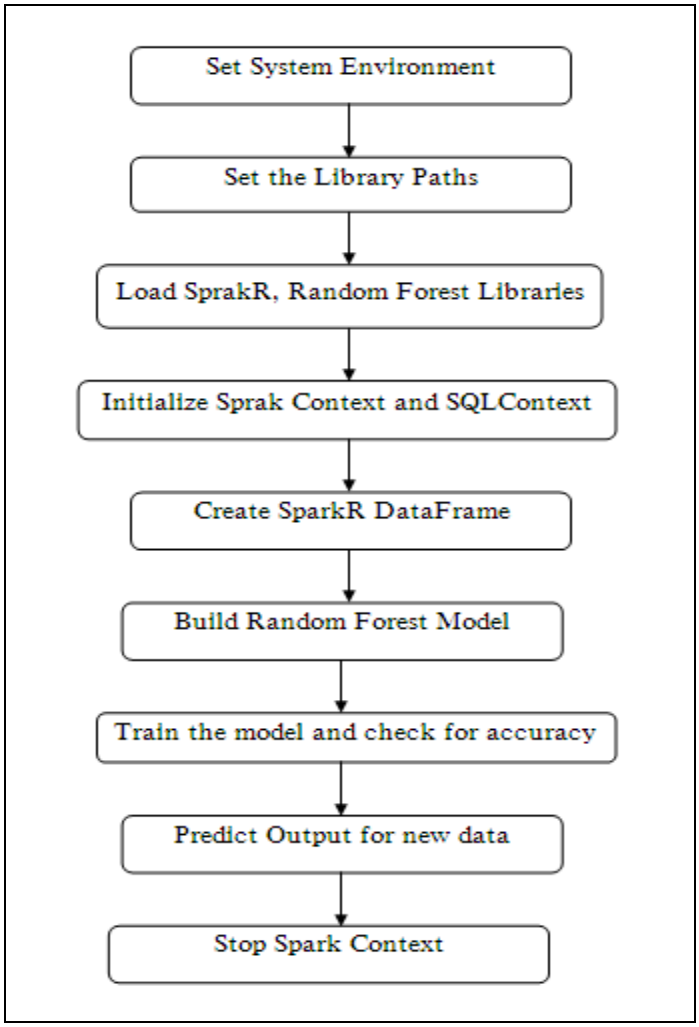

Fig 2: Control Flow Diagram of RF Modelling

Computational Details, Results and Discussions

This section explores particulars of experiment accomplished in SparkR and RStudio software platform for the classification of fetal states. RANDOMFOREST package in R environment is employed to study model structure [15]. We used the training data set for the parameter amendment of model whereas validation set to manage learning process. We tuned RF model with two parameters n_{tree} and n_{try} to get optimized forest architecture. The parameter n_{tree} specifies number of trees is to be built to populate the random forest where asn_{try} specifies the how many variables that will be considered in deciding partitioning of the dataset. We carried out performance evaluation for various RF configurations. Table 2 summarizes the experiment conducted per variation in $n_{tree}$by keepingn_{try} is 7 constant and shows performance of corresponding RF model. We have explored error plot and ROC curve as useful analytic tool for our random forest modeling. Error plot shown in figure 3 depicts optimal number of trees to build since the plot error rate gradually for the number of trees built.

Table 2: Performance evaluation for accuracy of Random forest Configurations

No. of Trees (n_{tree})	Mean Squared error	% Variable Explained
100	0.05454002	85.54
200	0.05339913	85.85
300	0.05368493	85.77
350	0.05421283	85.63

ISSN: 0975-8585

400	0.05231995	86.13
450	0.05327918	85.88
500	0.05292649	85.97
600	0.05265616	86.04
700	0.05243312	86.1

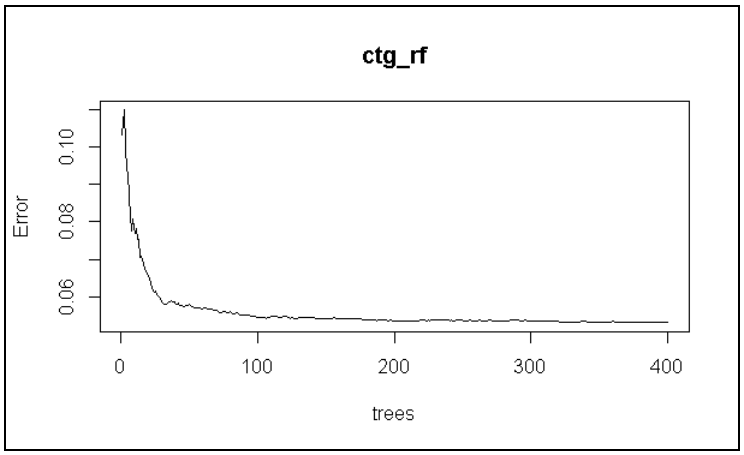

Fig 3: Error plot (MSE) for Random Forest model t

The optimized RF architecture chosen for the modeling of CTGs entails 400 trees in the forest with 7 partitioning variable. RF Model has used1488 observations randomly to build the forest. A textual representation of optimized RF model is given in fig. 4. The performance of RF modeling pertaining to this is shown in figure 5(a-b). In this case, mean square error (MSE) rate found to be 0.05231995. Further the overall assess of accuracy is then given by a confusion matrix that records the dissimilarity between the predictions and the actual outcomes of the training observations. We have tested model with known CTG samples. Fig. 6 shows confusion matrix for training dataset as well as for test dataset. Result concludes that RF modeling is a suitable approach since the resulting study is much more precise.

```
> print(ctg_rf)

Call:
 randomForest(formula = formula1, data = df, ntree = 400, mtry = 7)
               Type of random forest: regression
                     Number of trees: 400
No. of variables tried at each split: 7

          Mean of squared residuals: 0.05231995
                    % Var explained: 86.13
```

Fig 4: Textual representation of selected RF model

ISSN: 0975-8585

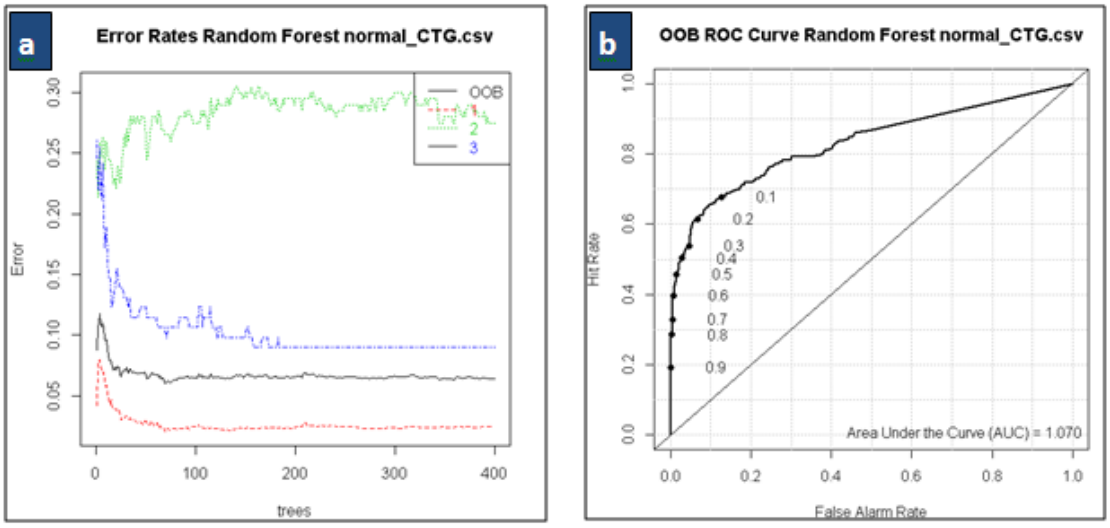

Fig 5: Performance of selected RF model with $n_{tree}$is 400 and n_{try} is 7; Fig(a) represents mean square error plot; Fig(b) represents ROC curve based Out-of-bag (OOB)

```
Confusion matrix:
     1   2   3 class.error
1 1137  21   8  0.02487136
2   50 145   5  0.27500000
3    6   5 111  0.09016393
```

```
          Predicted
Actual    1   2   3
     1  245   1   1
     2   16  29   0
     3    1   2  25
```

Fig 6: Confusion Matrix for training dataset and test dataset

CONCLUSION

This paper demonstrates machine learning approach for the modelling of fetal states by exploring Cardiotocography. The dataset with 2126observations of CTGs were selected for aforesaid investigation. In order to get optimum RF architecture, we have varied attributes such as number of trees and choice of variables for partitioning the dataset. The resulted RF architecture entails 400 trees in the forest with 7 partitioning variable. This model has chosen1488 observations randomly to construct the forest. Thus derived RF model efficiently classifies CTG samples into the given three fetal state classes with very less error. Thus the result recommends random forest has the potential to exhibit as the best tool for modeling of CTG samples.

REFERENCES

[1] Kamath RS, Kamat RK. Research journal of Pharmaceutical, Biological and Chemical Sciences 2016; 7(5): 2449-2455.
[2] Tomas P, Krohova J, Dohnalek P, Gajdos P. 36th International Conference Telecommunications and Signal Processing 2013; 620 – 923.
[3] Sahin H, Subasi A. Applied Soft Computing 2015; 33: 231-238.
[4] Sundar C, Chitradevi M, Geetharamani G. International Journal of Computer Applications 2012; 47(14): 19-25.
[5] Karabulut EM, Ibrikci T. Journal of Computer and Communications 2014; 2: 32-37.
[6] Magenes G, Signorini MG, Arduini D. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks 2000; 3: 637 – 641.
[7] Lichman M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California 2013; School of Information and Computer Science.

[8] Kamath R, Kamat R. Educational Data Mining with R and Rattle, River Publishers, Netherland, 2016, pp. 65-67.
[9] https://spark.apache.org/docs/latest/sparkr.html, Retrieved on 15th July 2016
[10] https://www.rstudio.com/, Retrieved on 15th July 2016
[11] Graham W. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer, UK, 2011, pp. 245-268.
[12] Grivell RM, Alfirevic Z, Gyte GM, Devane D. Cochrane Database Syst. Rev. 2010; 1.
[13] Alfirevic Z, Devane D, Gyte, Gillian ML. In Alfirevic, Zarko. Cochrane Database of Systematic Reviews 2006; doi:10.1002/14651858.CD006066.
[14] Kamath RS, Dongale TD, Pawar P, Kamat RK. Research journal of Pharmaceutical, Biological and Chemical Sciences 2016; 7(4): 830-836.
[15] Breiman L. Machine Learning 2001; 45(1): 5-32.
[16] Ocak H. J. Med. Syst. 2013; 37(2): 1-9.
[17] Menai MEB, Mohder FJ, Al-mutairi F. J. Med. Bioeng. 2013; 2(1).
[18] Huang M, Hsu Y. Journal of Biomedical Science and Engineering 2012; 5: 526-533.