An Approach for Data Analysis using Random Forest Algorithm

M. Mahaboob Basha[1], Dr. B. Lalitha[2] M.Tech[1], Assistant Professor[2]
CSE DEPT[1, 2], JNTUACEA[1, 2], Ananthapuramu[1,2], A.P, India[1,2]

Abstract

Data analytics is the way to organize and analyze the raw data, to make insights from the information. This uses latest techniques that are associated with data analytic processes in order to make conclusions with the help of automated algorithms. Based on an exploratory inspection of data modeling and from its understructure, a new practical approach is introduced to approximate multiple data properties, which is entirely based on the experimental examinations of distinct features and the distance between these features is termed as Empirical data analysis (EDA).The collective functions consists of, a measure of intimacy and distinctiveness. A unique property of the advanced practical approach for analyzing the data is, it doesn't depend on the random assumptions of the data. So, EDA is appropriate for the real situations like huge amounts of data is present. It is also applicable where it is not easy to find the hidden phenomena associated with the large sets of unclear data. Here we also show the appendage of EDA for assumptions. The proposed data analysis are widely applicable to solve real world problems. Even though it is implemented on the preliminary stages of data modeling, the initial examination results shows better performance compared to conventional Methods.

Index Terms — **Data Modeling and Science, Natural Language Processing (NLP), Statistics, Machine Learning (ML).**

I. Introduction

At present, data science is ruling the world with its systematic methods, algorithms to find out hidden patterns from ordered and unordered data. It is an interdisciplinary field associated with machine learning, probability and deep learning. Now-a-days data science related methodologies and the concepts are mainly focus on the real data, proof composed from practical examination results rather than theoretical preliminary inferences.

The primary element of the analytical approach is the definition of an arbitrary variable defines the probability law, i.e., a functional measure from the place of occurrence to the actual line. The calls of the conventional probability approach have a rigid base of mathematics and provide the ability to get high performance on data rich environment with same dissemination. Real data is normally arranged in discrete where traditional probability theory and statistics can be modeled as evidence of random variables, but the distributions about the data is unknown [19]. Good results can be possible if the hypothesis is confirmed about the pre data generation. If not there is a possibility to get a lot of failures. To overcome all these problems, the proposed system has a supervised model transformation of random decision forest tree with ensemble learning features. It is an ML based classification mathematical model that is composed of many decision trees to solve real life problems with high accuracy. It also uses Cross Validation to find out the generalization error [4]. The approximation of the error define the importance of training and test sets in evaluation of performance of the algorithm.

However, classification models are good

where the training data is very less and the output is binary classification. These classifiers will use less computational parameters like CPU, memory, and network. Overall, the proposed classification algorithms gives better performance within EDA framework [18] by randomly selecting training examples and random subsets of features for splitting of nodes to make decisions.

II. Related Work

ML [2] is also called as predictive analytics. It is defined as the ability to make systems learn and improve from experience without any explicit directions. It is mainly concentrating on developing the computer systems that can learn from patterns and inference techniques. In order to make decisions by using ML systems, the ML algorithms has built a mathematical model from sample data i.e., training data . ML is widely came into the picture where it is unable to develop the algorithm to perform the particular task. For ML, the mathematical optimization is important to deliver methodologies, theoretical and application domains.

ML is classified into categories based on their approaches, the input and output type of data and the problem they are going to solve.

Supervised Learning (SL):

SL algorithms build a model with a set of inputs and expected outputs. Here, the data is known as training data with a set of training examples. Machine only understands numbers so the mathematical model considers training as feature vector, represented by matrix. SL algorithms build a model that maps a function from input to output. An optimal model will make the function to predict the output with high accuracy on test data.

SL algorithms comprise of classification and regression. classification algorithms are nearly when the predictions are categorical values and regression algorithms are useful when the predicted output is the range of values.

P. P. Angelov et.al [1] proposed a naive EDA generalized methodology for discrete sets of non parametric models. The EDA is investigating unknown data models behind the data and opening up the framework for inference. For data analysis and to get the best results, considered similarity and distance metrics.

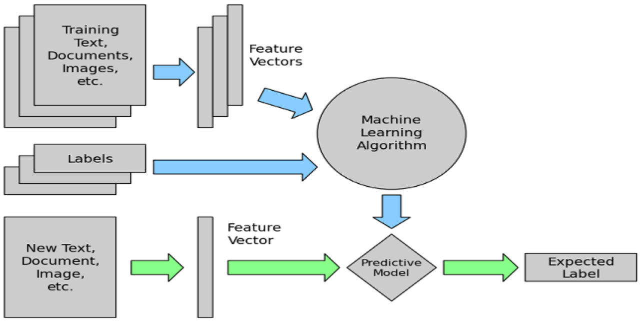
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Fig. 1. Supervised Learning Model

C. M. Bishop et.al [21] proposed pattern recognition(PR) and ML. It explains how the recent developments have been made by introducing them. These technologies useful to model the system and to classify the things with high efficiency.

Existing Work:

In the existing system Naive Bayes algorithm is implemented within empirical data analysis framework with less theoretical confidence levels for data modeling. Here, they considered only preliminary algorithms for classification.

Proposed Work:

As a proposed system, Random decision Forest algorithm has been introduced for data analysis. It gives high performance over preliminary algorithms and also it is implemented in different areas of applications.

It provides better results compared to other classifiers by implementing information gain.

III. Experimental Results:

A. System setup

The experiment was executed on each and every dataset to analyze the performance of the model. For every dataset we are taking some data to make the system to learn from its experiences. RDF trees also using same procedure to optimize the model performance. The performance is evaluated on a system with following requirements Intel® Core™ i5-3320M CPU@ 2.60 GHz, 64-bit OS and 8 GB RAM.

Random Decision Forest :

Here in this part of this section, we take the methodology of Random Forest (RF) [15] classification algorithm also called Random Decision Forest (RDF), a new version of random forest empirical data analysis. This new classifier depends on cross validation and correlation matrix [20] in contrast to original classifier.This approach is more effective in reflecting the collective properties of the dissemination of the different classes of data models in data space [5].

The proposed method accommodates multiple metrics like variance and bias [14]. To adjust the variance and bias, the user must have some knowledge so that it is easy to find the classifier for better classification.

Here is the list of known ingredients basis for the random forest i.e out-of-bag error (OOB) is used for estimation of the generalization error [4] and find out the importance of a variable by using permutation. Moreover, OOB estimate is the methodology to measure the prediction performance error of constructed RDF and the generalization error depends on the correlation and the trees strength. Most of the machine learning models uses bagging(Bootstrap aggregating) [3] to improve the stability, accuracy of the algorithm and to reduce the variance so that it eliminates overfitting.

Algorithm:

Precondition:

A training set T := (x1,y1)....(x_n, y_n), features M , and number of trees in the forest D.

```
function RandomDecisionTree(T , F )
    K←∅
    for i,... ,D do
N(i) ←A bootstrap sample from T
h_i←RandomForestLearn(N(i) , F)
K←K U {h_i}
    end for
    return K
end function

function RandomTreeLearn(T, M )
At each node:
    m ←very small subset of M
    Split on best feature in m
return learned decision tree
end function
```

RF classifier uses bagging methodology to train the model. Let us assume that the given training set is X and responses Y with bagging repetition.

for b=1..B

1. consider n training examples from X,Y with replacement is denoted as Xb,Yb.

2. with the training data, train a classifier fb on Xb,Yb.

After training, predict the performance of test data by averaging the predictions from all the individual trees on x' i.e majority of the votes.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Bagging decreases the variance without increasing the bias, it is a way of de-correlating the trees with different training sets.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

As we know, many trees are used to train the model based on the size and nature of training data. Cross validation(CV) is the way to find out the number of trees are required to train the model.

The performance of the proposed new RDF within EDA framework is tested on following real world problems:

1. Pima Indians Diabetes dataset [6].
2. SUV dataset [12].
3. Student records dataset [11].
4. SMS or Spambase dataset [10]
5. Heart disease dataset [7].
6. Employee dataset [9].
7. Credit card fraud dataset [8].

The proposed RF methodology is compared with logistic regression (LR) and k- nearest neighbor (KNN) to validate their performance on different problems of the real world [17].

In the proposed system, k-fold cross validation [13] is considered to train the learner to get more accurate estimate of the test data and more efficient use of every data. As mentioned, the performance is evaluated [16] after fitting the model with the data. Before validate the model it is necessary to train the estimator and then we test the model with test data because the model is already learn from the data.

B.Results:

Dataset	Overall Accuracy		
	RF classifier	LR classifier	KNN classifier
Pima	**78.64**	77.08	73.95
Student Records	**77.08**	70	57.50
Employee	**99.22**	77.86	74.82
SUV	**88.75**	62.25	74.25
Spam	**97.24**	95.51	95.63
Credit Card	**99.94**	99.915	99.916
Heart Disease	70.27	**77.86**	74.82

Table 1. classification performance with 5-fold cross validation

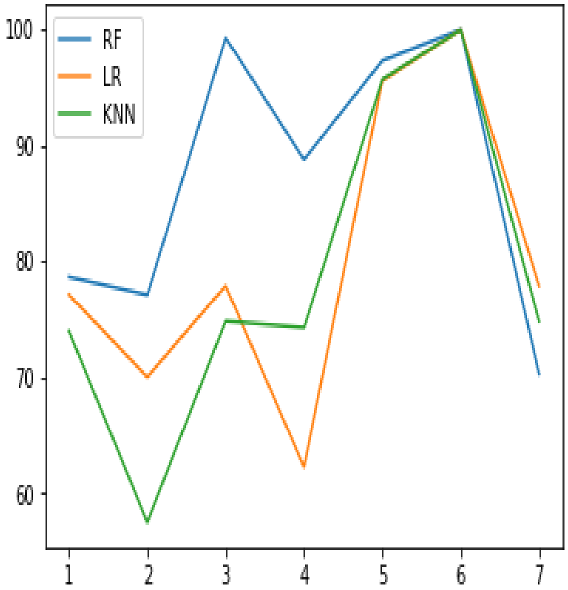

Fig. 2. Performance of models on multiple data sets with 5-fold cross validation

Dataset	Overall Accuracy		
	RF classifier	LR classifier	KNN classifier
Pima	**77.08**	72.69	72.65
Student Records	**75.69**	71.53	61.11
Employee	**99.30**	74.79	92.51
SUV	**89.50**	63.75	76.25
Spam	**97.24**	95.51	95.63
Credit Card	**99.94**	99.915	99.916
Heart Disease	75.88	**85.82**	77.82

Table 2. classification performance with 10-fold cross validation

In the above table, the performance of all the classifiers are noted with 5-fold cross validation is tabulated in Table I and Table II 10-fold cross validation are recorded.

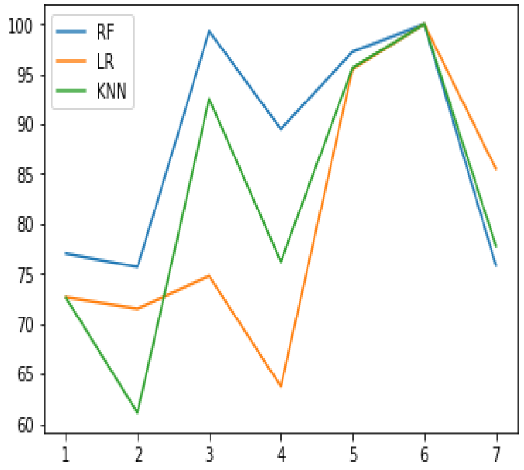

Fig. 3. Performance of models on multiple data sets with 10-fold cross validation

By seeing the results recorded in the above tables, the proposed RDF classifier gives better results over Logistic Regression and K-nearest neighbor classifiers on real world problems. Based on the validation results, the performance of the RDF tree model is the best. Finally it is shown that the proposed RF classifier is free from unrealistic assumptions about the data.

IV. Conclusion

In this paper, we propose a new well ordered proposal to obtain the collective properties of the facts without any advance information about the data sources, volume of data or specific boundaries regarding the data. A new classifier for the proposed system, RF to inquire into the unspecified data modeling hidden behind the vast amount of data in high data environment. In conclusion, the proposed RF approach within EDA gives an effective results on experimental and the confirmation. It is having base in many of the subjects i.e data mining and analysis, and therefore broadening the applications area, in particular, is in the big data and data mining.

However, we must acknowledge that the proposed algorithms have faith for the analysis of large amounts of data and with many number of features. Here, we provided supervised machine learning classification algorithms and its outcomes on data segmentation, investigation and categorization.

For future advancements in data analysis, we will concentrate more on unsupervised learning algorithms and deep learning algorithms to solve real world problems, including image classification and recommendation systems but not restricted to spam detection, disease prediction and fraud detection, etc.

REFERENCES

[1] P. Angelov, X. Gu, and Josh C. Principe, “A Generalized Methodology for Data Analysis,” IEEE., 2018

[2] Tom M. Mitchell, “Machine Learning,” 1997.

[3] Breiman, L. (1996a). Bagging predictors. *Machine Learning 26*(2), 123–140.

[4] Marianthi Markatou, Hong Tian, Shameek Biswas George and Hripcsak, “Analysis of Variance of Cross-Validation Estimators of the Generalization Error, ” Journal of Machine Learning Research 6 (2005) 1127–1168.

[5] Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844

[6] Pima Indians Diabetes Dataset. Accessed: Jul. 10, 2019. [Online]. Available:https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[7] Heart Disease Dataset. Accessed: Jul. 10, 2019. [Online]. Available:https://archive.ics.uci.edu/ml/datasets/ Heart+Disease

[8] Default of credit card clients Dataset. Accessed: Jul. 16, 2019.[Online]. Available:https://archive.ics.uci.edu/ml/datasets/ default+of+credit+card+clients

[9] Human Resource Dataset. Accessed: Jul. 20, 2019. [Online] Available:https://www.kaggle.com/rhuebner/hu man-resources-data-set

[10]Spambase Dataset. Accessed: Jul. 23, 2019. [Online] Available:https://archive.ics.uci.edu/ml/datasets/ spambase

[11]Student Performance Dataset. Accessed: Jul. 27, 2019. [Online]. Available:https://archive.ics.uci.edu/ml/datasets/ student+performance

[12]SUV Dataset. Accessed: Jul. 29, 2019. [Online] Available:https://www.kaggle.com/rakeshrau/soc ial-network-ads

[13]Payam Refaeilzadeh, Lei Tang and Huan Liu. "Cross-Validation." Springer, doi: https://doi.org/10.1007/978-0-387-39940-9_565

[14] Tibshirani, R. (1996). Bias, variance, and prediction error for classification rules. Technical Report, Statistics Department, University of Toronto.

[15] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[16] G. Santafe, I. Inza, and J. A. Lozano, “Dealing with the evaluation of supervised classification algorithms,” Artificial Intelligence Review, vol. 44, no. 4, pp. 467–508, 2015.

[17] S. Garcia and F. Herrera, “An extension on“statistical comparisons of classifiers over multiple data sets”for all pairwise comparisons,” Journal of Machine Learning Research, vol. 9, no. Dec, pp. 2677–2694, 2008.

[18] P. Angelov, “Outside the box: An alternative data analytics framework,” J. Autom. Mobile Robot. intell. Syst., vol. 8, no. 2, pp. 53–59, 2014.

[19] P. Angelov, X. Gu, and D. Kangin, “Empirical data analytics,” Int. J. Intell. Syst., 2017, doi: 10.1002/int.21899

[20]M. G. Kendall, “A new measure of rank correlation,”Biometrika, vol. 30, nos. 1–2, pp. 81–93, 1938.

[21] Stuart J. Russell, Peter Norvig (2010) *Artificial Intelligence: A Modern Approach, Third Edition*, Prentice Hall ISBN 9780136042594.