International Journal of Advanced Technology in Engineering and Science
Vol. No.6, Issue No. 05, May 2018
www.ijates.com
ijates
ISSN 2348 - 7550

THE EFFICIENCY OF RANDOM FOREST ALGORITHM IN BIG DATA ANALYTICS FOR MACHINE LEARNING

Dr. (Mrs) Ananthi Sheshasaayee [1] S.Malathi [2]

[1]Research Supervisor, [2]Research Scholar

PG & Research Department of Computer Science, Quaid -E- Millath Govt College for Women (Auto)

Anna Salai, Chennai-600 002, Tamil Nadu, India.

ABSTRACT

Random Forest algorithm is one of the most flexible and easy to use machine learning algorithm. At the same time it is also useful in analytics falling under various categories such as collecting of data, processing of data, analysing of data and finally interpreting of data Random forest algorithm holds good to achieve accurate data and also to find out the missing data. With growing volumes of data, faster and powerful computer, opportunity and use of analytics, many outcomes are concluded, among which machine learning have great and valuable impact. This paper aims to analyse such a compatibility between the random forest algorithm and the impact on big data analytics with reference to machine learning.

Keywords:*RFA Random Forest Algorithm), ML (Machine Learning), BDA (Big Data Analytics), RDF (Random Decision Forest), EDA (Educational Data Analytics), TA(Teaching Analytics), LA(Learning Analytics)*

I.RANDOM FOREST ALGORITHM

The most efficient and yet another algorithm for machine learning(ML)[14] is random forest algorithm(RFA). The random forest algorithm consist a collection of decision trees. Collection of several decision trees[21] are known as random forest. Classification technique uses many decision tree models that can be used to find out the accuracy, the missing data and other important information needed for decision making.RFA is well suited for predictive analysis.

Unless the other decision trees RF algorithm[4] does not have the limitations such as pruning as not all data are provided for decision making using trees. Instead only random subset of each data set makes the data set on the whole which is called as bootstrap/bagging[21].

In bootstrap process a data set contains sub-samples as well as attributes that are used for creating decision models. With data set support the classification or regression are used for decision model. Almost two third of data are repeated by every other decision tree. Using RF algorithm every decision tree outcome predicts a response for an instance and at the end, the final outcome is decided. In both regression and classification the

average of all the outcomes and the final outcome of the decision tree hold good for decision making respectively.

The RF algorithm is the most and popular algorithm for pattern recognition[13]. RFA provides aspects such as high and accuracy prediction[12], and information on variable that are important for data mining classification. The prediction performance of RF compares well to other classification algorithms.

The major features of RFA

1. RF is an effective algorithm for ML
2. RFA predicts accurate result even for large data base
3. RFA estimates the hidden data as well as missing data
4. In number of variables can be tested without major deletion process
5. RFA estimates the most important variables for classification technique
6. It holds good for prototype computation which is responsible for the relationship between the classification variable. It also computes the nearest pair of variables or rather test cases used for clustering technique.
7. RFA supports unlabeled data leading to clustering that are even unsupervised.
8. RFA aids detect the outlier and also data views

RF algorithm[5] is a popular method used to build models such as predictive models. These models can be used for estimating the outcome from both classification and regression problems. Several pattern[studies can be done based on random forest algorithm.

RFA advantages for ML

1. Maintains accuracy
2. Holds good for identifying missing data/ hidden data
3. Works well for large data set
4. Best suited for classification and regression technique
5. Holds good for model building as it reduces the dimension of attributes
6. Can also be used for unsupervised learning with unlabeled data

The impact of big data analytics in ML

BDA falls under much wanted category of today's world. BDA supports new opportunity in various fields. Aspects such as budget, flexibility and faster accessibility, decision making, creating new products and rendering services, model building falls under BDA. With the growing volumes of data, analytics and the impact in various fields such as business, marketing, customer service, operational efficiency, decision making, machine learning and also in competitive BDA application as well as in different data types forms a major

supports for predictive, decision and prototype models[11]. Machine Learning[10][15] algorithms are capable to learn from the relationships between predicted and actual outcomes.

BDA in technology

As there is no one fit technology that encompasses BDA there are several analytics applied to volumes of data which forms the most important factor for decision making[17].

1. **In data mining technology –** When using large amount of data which is used for discovering patterns which intern are used for complex question analysis used for decision making process.
2. **Used in-memory analytics** – Instead of referring from secondary memory, analyses can be done from system memory. By using this technology it removes data and analysis process inorder to test new scenarios and new models. These process helps to make better business decision.
3. **Used in predictive analytics –** For predictive analytics the data use the statistical algorithms and machine-learning techniques. The outcome is based on previous data and also provides with best future assessment[17].
4. **In text mining –**Text data used to analyze data from web. Text information from comment fields, hard copies like books and other text-based resources can be analyzed. Text mining mostly uses machine learning to comb through documents from emails, blogs, twitters, surveys, competitive intelligence in discovering new topics and relationships.
5. **In pattern mining** – Pattern mining helps in developing new courses in research areas.

BDA impact in EDA

BDA[18] and relevant application includes both internal and external data. In big data environment streaming analytics and applications are common[3].

Educational Data Analytics(EDA) Technologies

The meaning of data analytics refers to methods, tools involving large set of data and for analyzing data and to support in improving decision making. Current data analytics is also important in educational data mining which is very useful for higher education.

EDA technologies forming to be very useful to overcome real and practical difficulties for an effective data decision making in both teaching and learning. EDA[18] technology three main classification are:

Category 1 - **In teaching analytics(TA) – TA** refers to various methods and tools that involved in designing and developing a model for education[18].

Category 2 - **In learning analytics(LA) -** When referring to LA, it is the method of evaluation, data collection, data analyzation and summarization for optimized learning[18].

Category 3 - **In both teaching and learning analytics (TLA)** – TLA combines and supports category 1 and category 2 inorder to evaluate the learners and to support the ease of understanding the context of the resource material. For the teaching community it aids in improving the delivery of contents to the learners[18].

The compatibility between RFA and BDA in ML

RFA role of classification techniques are advantageous for

1. RF algorithm avoid the over fitting problem related to classification or regression and its application
2. RF algorithm used for identifying the important features from training dataset

Randomness presence is found in either one or other form in Machine Learning (ML) models. Especially the problem of over fitting, referring to a model that are meant as training data. Over fitting occurs when a model comes across the noisy detail in data set training to an extent of negative impacts performance of the model on new data.

The data set is split into two random sets for training. However change of paradigm for which randomness is the main accessory also plays a vital role in neural network area.

Since RF algorithm is more easy to use and flexible it produces more accurate values. EDA machine learning algorithm considers the huge set of data which is the further used for analysis. Using RFA model the technology predicts from the previous predictive results from decision trees which paves way to BDA.

II.CONCLUSION

RF algorithm can be used to predict newer and accurate outcomes. BDA also aids in organizations harness with the dataset thereby identifying many new opportunities. Newer techniques and technology provides various means of analyzing data with which many conclusions can be arrived. With growing data and RFA accurate results holds good for building various machine learning model which intern are useful for predicting future results benefitting both teaching and learning communities. Machine Learning algorithms that establish data-derived precedents incorporate them into future processes thereby proving predictive analytics enhancement in various analytics operations provide accurate data for better decision making.

REFERENCE

[1] Role of data mining in education field , Naik P.N , Volume 7, Issue 1, 2015, pp-215-217, 2015
[2] Big Data Analytics, (https://www.sas.com/en_in/insights/analytics/big-data-analytics.html)
[3] Big Data Analytics, Margaret Rouse, techtarget, March 2017
[4] Bagging and Random Forest Ensemble Algorithms for Machine Learning, Jason Brownlee, 2016
[5] Random Forests for Big Data , Robin Genuer, , Jean-Michel Poggi , Paris Descartes Christine Tuleau-Malot, Nathalie Villa-Vialaneix , 2017
[6] Participation in Higher Education: A Random Parameter Logit Approach with Policy Simulations Darragh Flannery, Cathal O'Donoghue,2009

[7] Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Wouter G. Touw Jumamurat R. Bayjanov Lex Overmars Lennart Backus Jos BoekhorstMichiel Wels Sacha A. F. T. van Hijum,*Briefings in Bioinformatics*, Volume 14, Issue 3, 1 May 2013, Pages 315–326, 10 July 2012

[8] Random forest algorithm in big data environment, Yingchun Liu ,September 2014

[9] Introduction to Random forest – Simplified, TAVISH SRIVASTAVA , analyticsvidhya, JUNE 10, 2014

[10] How data and analytics can improve education, Audrey Watters, July 25, 2011

[11] Pattern Recognition: The Paragon of Big Data Analytics, Shannon Kempe, 2014

[12] Improved Algorithm Based on Sequential Pattern Mining of Big Data Set, Peng Huang, Aug. 2016

[13] FREQUENT PATTERN ANALYSIS FOR DECISION MAKING IN BIG DATA, JULIJA PRAGARAUSKAITĖ, 2013

[14] Machine Learning With Random Forests, Anuj Saxena, dzone, Aug. 30, 17

[15]Machine Learning Algorithms Explained – Random Forests(http://blog.easysol.net/machine-learning-algorithms-2/), October 11, 2017

[16] Random forest algorithm in big data environment, Yingchun Liu, COMPUTER MODELLING & NEW TECHNOLOGIES 2014 18(12A) 147-151

[17] Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm, Julie Hardman, Alberto Paucar- Caceres ,Alan Fielding, Systems Research and Behavioral Science Syst. Res (2012) Published online in Wiley Online Library, August 2012

[18] Educational Data Analytics Technologies For Data-Driven Decision Making In Schools, Demetrios G. Sampson, October 20, 2016

[19] What Is Analytics? Team Jigsaw, jigsawacademy, Feb 2016

[20] Introduction to decision trees and random forests, Ned Horning , 2016

[21] Introduction to Random Forests, Raul Eulogio, Data Science, Aug 17