International Journal of Management, IT & Engineering
Vol. 8 Issue 4, April 2018,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

CLASSIFICATION AND OPTIMIZATION USING RF AND GENETIC ALGORITHM

Balaji K. Bodkhe*

Sanjay P. Sood**

Abstract

Data mining is process of extracting valuable and meaningful information from the large data set. There are many techniques but most important are classification and regression.In this paper we are discuss about classification algorithm i.e. Random Forest and Genetic Algorithm is for the optimization purpose. Random Forest is multi-classifier and it is used for increasing the accuracy of the classifier. RF constructs many decision trees from the given data set and it picks result of that classifier which is more accurate. In this paper we are combining both the algorithm i.e. Random Forest and Genetic Algorithm for more optimized result and to increase the accuracy of the classifier. So the class predicated by the different classifier in Random forest is provided to the genetic algorithm and then genetic algorithm gives the most accurate result from the given predicated result.

.

Keywords:

Big Data; Random Forest; Genetic Algorithm; De-identification policies

* Research Scholar, IKG Punjab Technical University, Kapurthala, Punjab, India

** CDAC, Mohali, India

1. Introduction

Big data term can be defined by the 3 V's Volume, Velocity, and Variety. Volume means the data collected from the different sources, Velocity means data should be processed in faster way, and Variety means data which comes from all type's structured, unstructured, text, audio, video etc. so to process such a data which is large in size and which has a variety we use data mining techniques. Big data is the collection of large dataset that cannot be processed by the traditional techniques and data produced by different devices and applications. This includes system likes massively parallel processing and map reduce that provide analysis of large dataset. Map reduce provides methods for analyzing data that is complementary to the SQL. In this paper we are using two algorithms i.e. Random Forest and Genetic algorithm. Random forest is ensemble learning method for classification, regression and other tasks that constructing multiple decision trees on health care dataset and it picks result of that classifier which is more accurate. Classification uses majority votes and regression uses average or mean. We are classifying health care dataset predicting disease which patient has and allocating doctor for that particular disease.

In Genetic algorithm is search based optimization technique based on principle of genetic and natural selection. Frequently used to find optimal solution is difficult problems. Optimization refers to finding value of inputs in such a way that it gets the best output values. These algorithm uses three basic genetic operators namely reproduction, crossover and mutation along with fitness function to evolve new population or the next generation. The significance of the genetic operators are reproduction or selection by two parent chromosomes is done based on their fitness, crossover or recombination is for making new chromosomes that possess characteristics similar to both the parents and mutation use to avoid local optimum.

2. LITERATURE SURVEY

1) A Comparative Study on Decision Tree and Random Forest Using R Tool

Authors: Prajwala T R

In this paper the author has compare two algorithms Decision tree and random forest using the R tool. Decision tree is classifier used to produce the decision tree. Decision tree has the decision node that specifies a test on an attribute, leaf node value of the target attribute, edge split one attributes. Decision tree is easy for analysis of numeric and categorical data. Decision tree is efficient to find outliers in dataset. Random forest is ensemble of decision tree method and multi-

classifier used to construct many decision trees and helps in predicting data accurately. At the end redistribution error rate of random forest is less than decision tree but time taken by random forest to execute dataset is more

2) Random Forest: **A Review**

Authors: EeshaGoel, Er. Abhilasha

This paper is all about random forest algorithm and where we can use the random forest algorithm its applications. Random forest is ensemble classification technique this is used to improve accuracy and performance of the classifier but on other hand it is time consuming as compare to other techniques. Random forest can be used in prediction of pathologic complete response in breast cancer, cause of death prediction, on-line learning and tracking etc.

3) Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey

Authors: Vanaja, S. and K. Rameshkumar

Aim of this paper is to study different classification algorithm used in medical dataset and compare its performance. The algorithms with the highest accuracy on various kinds of medical dataset are taken for performance analysis and the analysis shows most frequently used algorithm on medical dataset. In this paper C4.5 algorithm gives efficient performance than any other classification techniques and it can handle large amount of data, reduce error rate and gives better result for medical dataset.

4) Efficient Recommendation of De-identification Policies using MapReduce

Authors: Xiaofeng Ding, Li Wang, Zhiyuan Shao, and Hai Jin

In this paper we study de- identification policies using Map-reduce. Firstly policy generation of proposed definition which decreases time and size of alternative policy set. Secondly Sky-Filter-MR which is map reduce based parallel algorithm used to represent policy of the framework. Integrating skyline with map reduce algorithm, filtering power of map phase is optimized and also performance, scalability of Sky-Filter-MR is increased.

5) Genetic Algorithms using Hadoop MapReduce

Authors: RakeshYadav, Dr. Ashok Kumar Turuk

In this paper genetic algorithm is used in hadoop map reduce framework. Map reduce has the two keys map and reducer. Map actually maps the data form key and value then output of map phase data goes to intermediate phase and data is divided into cluster. Number of clusters depends on number of reducers and then reducer gives the result. Genetic algorithm is the search based algorithm and it used for optimization purpose. It has basic genetic operator reproduction, crossover, mutation and fitness function. According to Fitness value chromosome are selected.

3. RANDOM FOREST ALGORITM

This algorithm is also known as a multi classifier .It is an ensemble method, ensemble is a divide and conquer approach for example Group of weak learner come together to form a strong learner. This algorithm includes decision trees construction from the given training data set and matching the data with these, so it helps to analyze and predict data accurately. Random forest works for both classification and regression in that classification uses majority votes and regression uses average or mean.

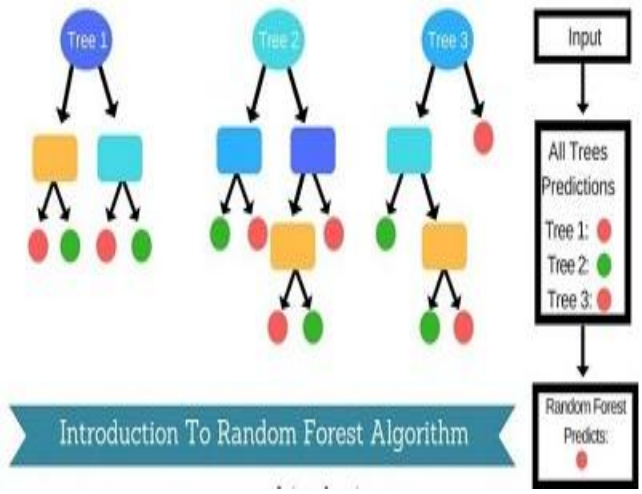

Fig. Random Forest

In our system there has database of diseases and doctor with the help of symptoms system predict the accurate disease and allocate particular doctor for that disease from the decision trees of diseases system take the majority vote to classify the disease and allocate specialist doctor.

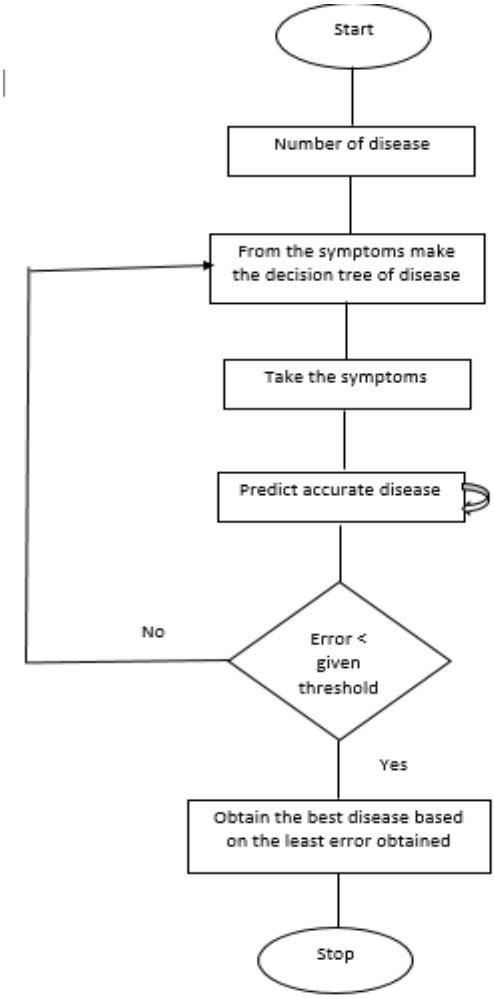

Fig. Diagrammatic Representation of Random Forest

Advantages:

1. It is robust to correlated predictors.
2. It is used to solve both regression and classification problems.
3. Runs efficiently on large databases
4. Requires almost no input preparation
5. Performs implicit feature selection
6. Can be easily grown in parallel
7. Methods for balancing error in unbalanced data sets

Disadvantages:

1. It takes care of missing data internally in an effective manner.

Applications:

1. Banking
2. Medicine
3. Stock Market
4. E-commerce

4. GENETIC ALGORITHM

The genetic algorithm is a search based optimization techniques. In GA's we have pool or population of possible solution to given problem for solving the particular problem .we are having 'n' number of solution but all the solution cannot provide optimize solution to that problem but the GA helps in solving all kinds of problem where it is constraint or unconstraint one.

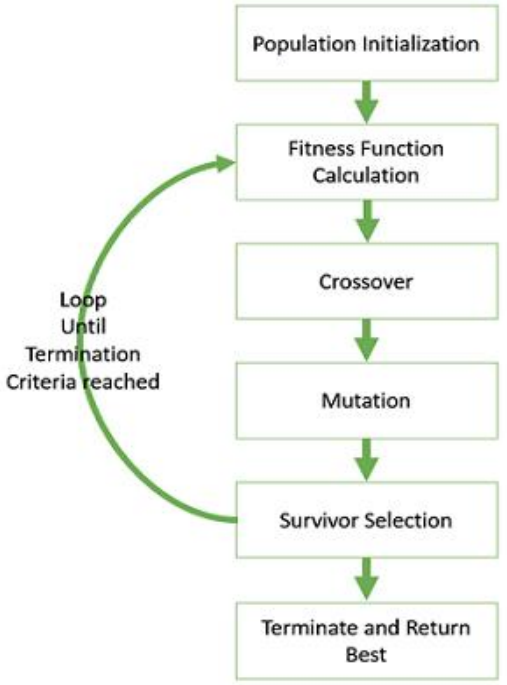

Fig. Diagrammatic Flow of Genetic Algorithm

In GA's, population is nothing but subset of all the possible (encoded) solutions to the given problem. And the chromosomes is one of such solution to given problem. The chromosomes is the bunch of genes. Most important term is introduced that is fitness function. A fitness function

simply defined is a function which takes the solution as input and produces the suitability of the solution as the output.

Steps

1. Choose a coding to represent problem parameters, a selection operator, a crossover operator, mutation operator, population size, crossover probability and mutation probability.
2. Initialize random population of strings of size l, $tmax_{, set}$ t=0.
3. Evaluate each string in population
4. if $t>t_{max}$or other termination criteria is satisfied, terminate
5. Perform crossover on random pairs of string
6. Perform mutation on every string
7. Evaluate strings in the new population. Set t =t +1 and go to step 3

6. GOALS AND OBJECTIVE

- The main goal of the project is to study, design and implement performance optimizations for big data frameworks. This work contributes methods and techniques to build tools for easy and efficient processing of very large data sets. It describes ways to make systems faster, by inventing ways to shorten job completion times.
- To generate faster results.
- It reduces the complexity of data access and retrieval. When we have to dealing with big data.
- The alternative to this is apache Hadoop, which deals with big data with efficiency.
- Hadoop itself consists of Map Reduce and HDFS.
- Provide security to personal information.
- Protect the user data during transmission.
- We perform a detailed security analysis and performance evaluation of the proposed technique.

7. CONCLUSION AND FUTURE SCOPE

We study the recommendation on a great number of de identification policies using Map Reduce. We can handle easily large amount of health care dataset with the help of Hadoop and map

reduce. HDFS to store large amount of data and map reduce is for to process on that data. We can easily classify and optimize data with the help of Random Forest Algorithm we can easily classify the health care data and Genetic Algorithm to optimize that classified data. This algorithm is for accuracy, low error rate and high performance. This web application is easily can handle the patient, doctor, chief doctor and admin.

References

[1] *Prajwala T R, "A Comparative Study on Decision Tree and Random Forest Using R Tool", in ISSN, Vol 4, pp. 1-4, Jan 2015.*

[2] *Eesha Goel, Er. Abhilasha, "Random Forest: A Review", in ISSN, Vol 7, pp. 1-7, Jan 2017.*

[3] *Vanaja, S., K. Ramesh Kumar, "Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey", in J of CS, pp. 1-23, Jul 2014.*

[4] *Xiao Feng Ding, Li Wang, Zhiyuan Shao, and Hai Jin, "Efficient Recommendation of De-identification Policies using MapReduce", in IEEE, pp. 1-6, 2016.*

[5] *Rakesh Yadav , Ashok Kumar Turuk, "Genetic Algorithms using Hadoop MapReduce", in NIT Rourkela, pp. 1-36, 2015.*

[6] *K. Benitez, G. Loukides, and B. Malin, "Beyond safe harbor: Automatic discovery of health information de-identification policy alternatives," in IHI, 2010, pp. 163–172.*

[7] *K. E. Emam, "Heuristics for de-identifying health data," IEEE Security and Privacy, vol. 6, no. 4, pp. 58–61, 2008.*

[8] *W. Xia, R. Heatherly, X. Ding, J. Li, and B. A. Malin, "Ru policy frontiers for health data de-identification," Journal of the American Medical Informatics Association, vol. 22, no. 5, pp. 1029–1041, 2015*