VOL. 12, NO. 15, AUGUST 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences
©2006-2017 Asian Research Publishing Network (ARPN). All rights reserved.
www.arpnjournals.com

HEALTH ANALYSIS USING BIG DATA

Sandhva P.
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai,Tami Nadu, India
E-Mail: sandhya.p@vit.ac.in

ABSTRACT

The massive amount of data is being extracted in every day's life. Some data is very useful and some is just the garbage that means data of no use. A key term came to existence while working with or handling large data that is BIG DATA. Big Data is a processing system which computes the large data, analysis it and predict the meaningful outcome. The major role of Big Data is being played in the health care industry. If we carefully analyze the health care industry one of the most dangerous diseases is"Cancer". This Paper focuses on Breast Cancer a major problem that has been increased in numbers. To analyze the risk of a patient there are number of factor's involved and these factors make it tedious process which is hard to analyze. To simplify the process we will analyze the data set and apply the machine learning algorithm. The data set will be simplified and using R-tool we will implement the random forest algorithm.

Keywords: breast cancer, risk factors, big data, R-Tool, hadoop, random forest algorithm.

1. INTRODUCTION

Data is generally a fact and figures, to extract meaning from the data is known as information. These data are collected and retrieved from various sources in massive volumes, high rates and different structures example: data collected from space, social sites etc with these characteristic is called Big Data. Working with big data is quite a big challenge. Its about taking input, processing and getting output. To work with a large data there are so many algorithms that need to be implemented. A Map reduce framework was build as a parallel distributed programming model to process such large scale of data set effectively and efficiently. So large data set will be taken and algorithm will be implemented and an output will predict the knowledge.

A numerous cases have been reported for breast cancer patient. A breast cancer patient has so many factors that causes these disease. In this paper the cell nuclei is of breast mass is taken into existence, which have various factors texture, radius, smoothness etc. On this basis the identification is done whether the cell is Benign, Malignant.

Benign means the tissues that don't spread among other tissues where as Malignant means the growth of tissues is spread among other tissues. To identify this for given data set we will conclude by using the decision tree algorithm on the given data set. Breast Cancer is cancer that develops from breast tissue. Its also one of the most dangerous diseases in our era. About 246,660 case have been estimated to be spreada and the number is increasing.

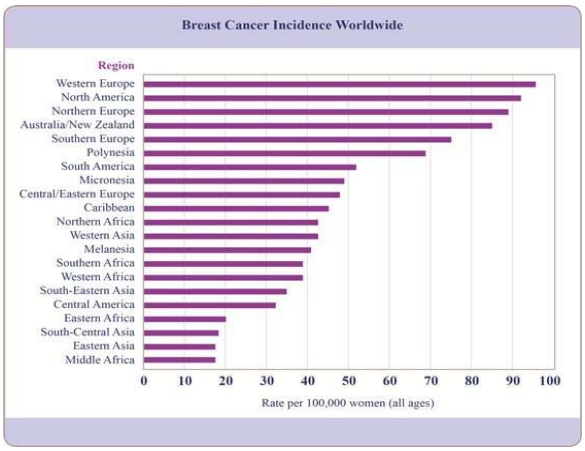


Figure-1. Breast cancer graph.

The following graph will show the number of reports of breast cancer form various region.

2. RELATED WORKS

a) Rostom Mennour and Mohamed Batouche presented their work on breast cancer analysis using big data and Hadoop in this paper the virtual screening technique is used for identifying the drug discovery using Mahout and Map Reduce.

b) Dr. VrindaTotekar and Shweta Pandey presented the work on Big Data using map reduce. Key role is being played by the Big Data in industries for betterment business decisions but the main issue with Big Data is, its analysis, MR is playing a significant role in Big Data analysis and connecting to the edge of parallel DBMS. The performance of MR can improve the tuning quality but still the research work is required in the same direction.

c) Karamjit Kaur Boticario Application of Data Mining for High Accuracy Prediction of Breast Tissue Biopsy Results planned an Accessible and Adaptive module, the Mammogra-phy Masses data set are highly

©2006-2017 Asian Research Publishing Network (ARPN). All rights reserved.

promising and have an accuracy of 83.5 percentage.

d) E.S. Samundeeswari and P.K. Saranya The PRI, VoI, GCE and BDE cluster validation measures are used to compare

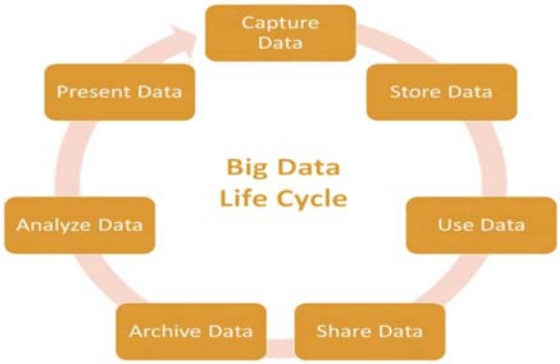


Figure-2. Big data life cycle.

the segmented result with ground truth image delineated by the radiologist. The proposed work outperforms the traditional K-Means clustering method with 96 similarities (PRI) between segmented tumor images with referred tumor image.

3. BACKGROUND

As per the introduction it has been mentioned there are few terminology and software to accomplish our objective. We will present them in this paper for better understanding.

A. Map reduce/Hadoop

In the Big Data Community, Map Reduce has been considered as one of the main methods to meet constantly increasing demand on IT resources imposed by massive data sets. The reason is the high scalability of the map reduce that enables massively.

The Map Reduce system is presented as a straight forward and effective programming model that empowers simple ad-van cement of adaptable parallel applications to handle inconceivable measures of information on expansive bunches of product machines. The best of Map Reduce is that the developer does not need to stress over the usage elements of parallelism and adaptation to internal failure, the framework deals with it for him. The software engineer just needs to consider how to adjust the issue to the model. To do as such, he should characterize two obligatory and essential capacities: the Map () work and the Reduce () work. Hadoop is the most celebrated execution of the Map Reduce Model, it's an open source preparing framework for enormous information, and it utilizes a circulated record framework i.e. HDFS (Hadoop Distributed File System) as a capacity stage. Hadoop is exceptionally reasonable for cluster sort issues, and it can be intense and versatile with regards to handle and examine vast amounts of information. Likewise Hadoop is a blame tolerant framework, which implies that regardless of the possibility that one or numerous hubs in the group crash, the framework proceeds with its execution with no issue. That is the reason we have picked this framework to play out our encounters.

B. R-Tool

R is a programming tool and software for representation and measurable for registering. R is widely used for statically purpose and data miners for data analysis is an open source software and operating system independent. In R language there are predefined libraries and which consist of predefined functions. With the help of it library function there is various advantages of r as per graphical representation there are various end to end graphics.

R is an interpreted language. Generally the user uses it by using command line. Like several other languages the r has its own libraries functions. Which are used for various purpose. It is an easy and simple to implement language.

4. PROPOSED WORK

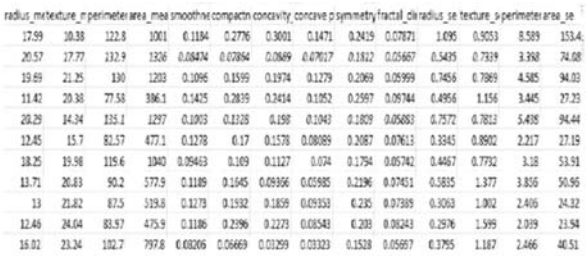
In the past work, segment will clarify how to build up the proposed approach for medication revelation with regards to breast malignancy. To start with we will depict the data set and how to infer it. And after that, we will exhibit the picked calculations and methodologies and how we utilize them to finish our objective.

Dataset-In this work the advantage is centered around the breast tumor protein, so to choose the receptor 4JLU which is a precious stone structure of BRCA1, this protein is accessible in the protein information bank (PDB). Since this receptor has not been investigated some time recently, it has been obliged to build the dataset from the scratch. It contains 106 distinct ligands in the PDBQT organize. 104 ligands were haphazardly chosen to build dataset, and after that the virtual screening procedure was performed utilizing Auto dockVina. To have a superior thought of the variety of results, a diagram was attracted to demonstrate the recurrence of estimation of partiality (Kcal/mol) that we have, and the chart has been plotted. We take the middle as a partition point, which is - 5.8 Kcal/mol for the situation. The partition indicate is utilized separate amongst dynamic and dormant ligands, and it's up to the analyst to choose its esteem, selecting a high esteem implies that it will be a significant number of genuine positive that will be ignored, on the creator hand, selecting a lower esteem implies the expansion of false positive number, that is the reason to pick the middle. The Molecular configuration PDBQT is known to be hard to comprehend and to break down, so in this work we have changed over every one of the ligands to the unique mark organize (FPT) which is more advantageous to be utilized as a part of machine learning calculations. Open Babel, an open source concoction tool compartment that permit transformations between compound structure configurations, was utilized to play out our change. The FPT string was in hexadecimal, so further change to two fold, the subsequent dataset was a n * m framework, where n = 104 is the quantity of preparing cases, and m =

©2006-2017 Asian Research Publishing Network (ARPN). All rights reserved.

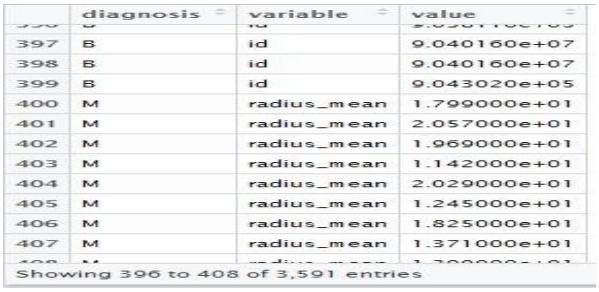
1024 speaks to the quantity of double components. A vector of 104 components was likewise made and imported to speak to the mark class into the information set.

Models-After the information development stage is done; the dataset was utilized to prepare five models. Calculation intended to be utilized as a part of the setting of huge information on top of the stage Hadoop/MapReduce were utilized as a part of this work, we are speaking here about calculations like the versatile irregular woods calculation in view of MapReduce, and the mahout innocent execution and so forth. Every one of the calculations that we have utilized here are actualized by MapReduce demonstrate on the Hadoop stage utilizing the java dialect. In this work we have utilized mahout to fulfill our examination, then we have done a correlation between the calculations and we have chosen the best three of them to construct a group of classifiers



radius_m	texture_m	perimeter	area_mea	smoothne	compactn	concavity_	concave p	symmetry	fractal_dia	radius_se	texture_s	perimeter	area_se	
17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	
20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	
19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	
11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	
20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	
12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	
18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	
13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	
13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	
12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	
16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	

Figure-3. Sample data set.



	diagnosis	variable	value
397	B	id	9.040160e+07
398	B	id	9.040160e+07
399	B	id	9.043020e+05
400	M	radius_mean	1.799000e+01
401	M	radius_mean	2.057000e+01
402	M	radius_mean	1.969000e+01
403	M	radius_mean	1.142000e+01
404	M	radius_mean	2.029000e+01
405	M	radius_mean	1.245000e+01
406	M	radius_mean	1.825000e+01
407	M	radius_mean	1.371000e+01

Showing 396 to 408 of 3,591 entries

Figure-4. Output of data set after using melt function.

5. IMPLEMENTATION PLAN

In this paper we will be working with data set that has been taken from the Kaggle website. The data set explains about the various factors of cell nuclei just about the mean of smoothness, texture, concave etc.

This data set will be taken and using r tool first it will be preprocessed and using melt function in r tool. This help to identify the filtered data from what we got. Generally all the function have predefined libraries. The libraries contain method definition and structure. For example the Re Shape 2 library contains a melt() function.

After getting the filtered data the random forest algorithm will be applied over the data set to identify which part of cell is affected at very first and what are the factor that is concluded in the cell.

Using filtered data the random forest algorithm will provide decision Description: To extract which predictor variables will predict the output of random variable. sample.ind ¡- sample(2, nrow(r), replace = T, prob = c(0.6,0.4))

varName ¡- varName[!varName

varName1 ¡- paste(varName, collapse = ”+”)

rf.form ¡- as.formula(paste(”y”, varName1, sep = ” ”))

cross.sell.rf ¡- randomForest(rf.form,b.s.v,ntree=500,importance=T)

varUsed(randomForest(Species ., iris, ntree=100)).

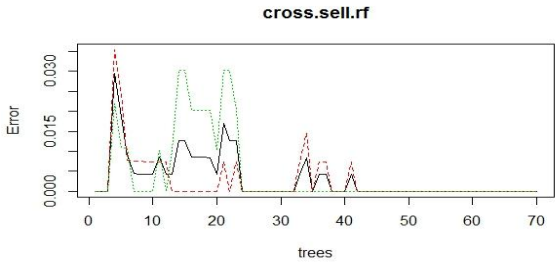


Figure-5. Output.

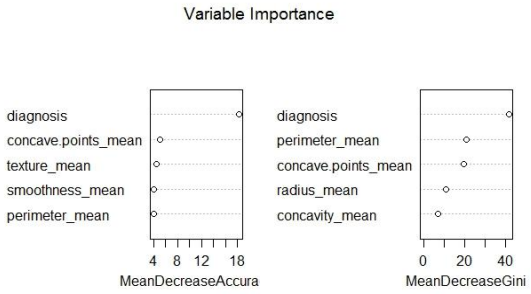


Figure-6. Plotting.

6. CONCLUSION AND FUTURE WORK

In this paper we come to conclusion that how the cell has affected the most dangerous disease Breast Cancer. In this paper we worked with the factors of cell nuclei.

In this paper the Key idea was to work with the cell that identifies the Benign or Malignant.

Benign means the tissues that don’t spread among other tissues where as Malignant means the growth of tissues is spread among other tissues.

This paper concentrates on ”Breast Cancer” a one of the most hazardous sickness on the planet. To examinations the danger of a patient to get breast malignancy various variables must be represented wish turns into an exceptionally repetitive process; this is because of the way that such a large number of factors make it hard for us to investigation. To rearrange this

©2006-2017 Asian Research Publishing Network (ARPN). All rights reserved.

methodology we play out the multivariate examination of the breast growth chance variable information set from the breast disease.

We have learnt machine learning calculations intended for huge information examination to do the characterization of cell cores into dockable and non dockable ones. After that, we have chosen calculations which are the best ones in term of exactness as per our involvement with the breast tumor receptor, and we have made a group of classifiers utilizing these calculations.

As we trust that more information will enhance the model we planned, we will move it to greater bunch of machines where we will have the capacity to perform it on a higher number of ligands in a generally better execution time. Addi-tionally we would like to investigate all the more enormous information calculations for machine learning with regards to virtual screening

ACKNOWLEDGEMENT

I acknowledge Ripudaman Singh Rathore, student, SCSE for his contribution in the paper.

REFERENCES

[1] Rostom Mennour: Drug Discovery for Breast Cancer Based on Big Data Analytics Techniques.

[2] Dr. Vrinda Totekar: proposed Prominence of Map Re-duce in BIG DATA Big Data is playing a very significant role in industries for making better business decisions but the main issue with Big Data.

[3] Karamjit Kaur Boticario Application of Data Mining for High Accuracy Prediction of Breast Tissue Biopsy Results planned.

[4] E.S. Samundeeswari P. K. Saranya Department of Com-puter Science, R. Manavalan.

[5] The Apache Mahout Project: http://mahout.apache.org/.

[6] J. Dean and S. Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. Commun ACM. 51(1): 107-113.

[7] Sherif Sakr and Anna Liu, Nicta and Uni-versity of New South Wales AYMAN G. FAYOUMI, King Abdulaziz University, The Family of MapReduce and Large-Scale Data Processing Systems, ACM Comput. Surv. 46, 1, Article 11 (October 2013).