PREDICTIVE BIG DATA ANALYTICS USING EVOLUTIONARY COMPUTING FOR CANCER DETECTION –A REVIEW

Dr. Alamelumangai. N

Abstract

A patients' record generates large volume of data; if the data are managed and analyzed many solutions and patterns to problems can be identified leading to diagnose of cancer. This will help the doctors to take proper decisions. Predictive big data analytics is such a process of finding a statistical technique to predict, model and analyze the information. In this system, structured meaningful data are mined using clustering technique. Memetic Algorithm is combined with analyzed text big data from electronic medical records and integrates them to detect the occurrence of cancer. The system uses data mining techniques, statistical analysis, machine learning, modeling and artificial intelligence to analyze and predict the future. Predictive analytics model is used to hold the relationship between the factors to assess risks among cancer big data and assign score or weights. This paper presents a review of predictive big data analytics using evolutionary computing an indicative way out to humanity for early detection and prevention of cancer.

Keywords: Predictive Analytics, Big Data, Evolutionary Computing, Cancer Detection.

I. INTRODUCTION

The amount of medical data generated every day is expanding in drastic manner. These data are stored in data warehouses causing as they are in raw format; proper analysis and processing is to be done in order to produce usable information out of it. To analyze these huge data - classification, cluster analysis, crowd sourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, predictive modeling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualization are applied. Big data is a popular representation to describe the data in zetta bytes. Additional technologies used to analyze big data are- massively parallel-processing (MPP) databases, search-based applications, data-mining grids, distributed file systems, distributed databases, cloud computing platforms, the Internet, and scalable storage systems.

Big data has few key characteristics such as volume, sources, velocity, variety and veracity [1]. Big data is currently a recent topic and it uses the technologies like Map Reduce, Hadoop, etc. Data is a combination of both structured and unstructured data.

Predictive analytics is the study of which is used to make predictions about future events. Data mining, text mining analytics along with statistics are used to predict acute patterns and relations structured cancer data [2]. Textual data is extracted from the text. Predictive model is used to describe the relation chain between the performance of a unit in a sample space and another known feature. The main goal of the model is to cluster the cancerous data. Predictive models iteratively evaluate the transactions to diagnose the occurrence of cancer cells [5].

II. PROBLEM AND OBJECTIVES

Analyzing and managing big data related to different Cancer Patients worldwide and generating a prototype which will be implemented on Hadoop which will help the doctors in Diagnosis and Prognosis of the Cancer Patients. Designing a Prototype using HDFS (Fig.1) which will effectively collect, clean, analyze the data from different data sources. Developing an algorithm can produce accurate results for the better Decision making

in Diagnosis and Prognosis of the Cancer Patients. The main objective of this paper is to help the doctors in saving the lives of millions of people who are suffering from different types of Cancers worldwide. Big data is used to unearth the hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informative decisions [4].

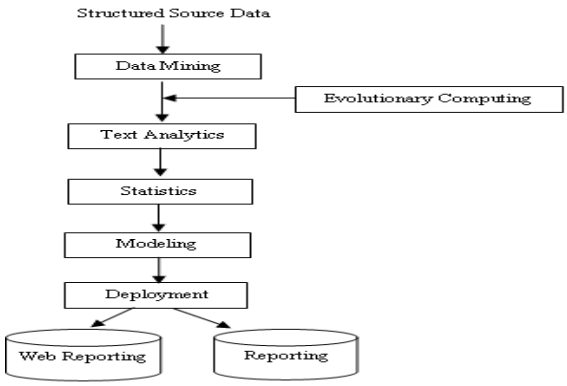

Fig 1. Flow of Predictive Analytics

III. MEMETIC ALGORITHM

Memetic Algorithms (MAs) are adaptive heuristic investigative algorithms based on the natural selection and genetics. They represent an intelligent, explorative and random search to solve optimization problems. MA explores the characters of the error functions. MA uses the mechanism of natural selection and genetics to its population of solutions. It involves - Global optimization, stochastic searching and Selection is based on good features. The general features of MA which make it suitable for the likelihood of choosing the operators is given in Figure 2.

```
Procedure
Begin
  Generate the queen string
  Initialize the bit strings          Fitness function
  While 100 generations
  Begin
      Evaluate the child string
      Choose mutation points          cross over
      For each string
      Begin
         Decode parameter             mutation
         Assign Filter
      End for
   End while
End procedure
```

Figure 2. MA Algorithm for Parameters Optimization

The initial population which is chosen by the MA algorithms is the one individual string which is defined as the 'queen' string. The generations are generated by conducting mutation operations on these strings. These strings contain the membership function (msf) width-parameter P and the threshold weights have to be applied between the input and hidden layers.

IV. HDFS IMPLEMENTATION

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. Applications that run on HDFS have large data sets. A typical file in HDFS is gigabytes to terabytes in size. The algorithm is applied on Bigdata(Cancer Patients Data provieded by National Institute of Cancer) by using the HDFS file system and it follows as:

i. Randomly initialize the structured Cancer data
ii. Determine fitness
iii. Repeat the tasks
    a. Select parents from population
    b. Perform single point crossover on parents
    c. Perform mutation of population
    d. Determine the fitness of population until best individual is good enough.

The experiment is conducted on Ultrasound images which are trained analysed statistically and modelled for deployment in HADOOP environment [11]. The input parameters are adjusted for a feasible result. The system performance is tested of its error value based on Mean Square Error (MSE) [12]. The results are compared with the simulations of the existing models adaptive mean filter. It is observed that the noise from the image is reduced considerably. The experiment is simulated [13] and compared with existing models (Table I) using *Rapid Miner* tool on Athol processor based system with 2 GB RAM.

Table I Comparison of Noise Mean Square Error

Method	***EPOCS***	
	100	**200**
Adaptive Mean Filter	0.529	0.512
MA Based Method	0.492	0.478

V. CONCLUSION

Intelligent systems have been applied to the diagnosis of different diseases and in various fields. As and more data move online, we are compelled to see many predictive solutions from the monitoring of data to the detection of fraud and abuse. The usage of MAs does not guarantee the parameter space. All the tools become precise in the availability of large volumes of digital data. Usage of various predictive analysis tools in Big Data HADOOP environment helps find the future cause. There is a necessary to choose correct tools based on the applications various significant factors in Big Data environment.

REFERENCES

1. Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review", Retrieved 7 October 2015.
2. Bernard Marr, “Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions”, 2015.
3. Bernard Marr, “Big Data for Small Business For Dummies”, November 2015.
4. T. White, “Hadoop: The Definitive Guide” , Third ed., O'Reilly Media, Yahoo Press, 2012.
5. Eric Siegel, “Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, Or Die”, 2013.
6. Dr. Anasse Bari, Mohamed Chaouchi, Tommy Jung, “Predictive Analytics For Dummies”, 2013.
7. Bala Deshpande, Vijay Kotu, “Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner”, 2014.
8. Keith McCormick, Jesus Salcedo, “SPSS Statistics for Data Analysis and Visualization”, 2017.
9. The Forreater Wave, “Big Data Predictive Analytics Solutions”, 2013.
10. Daniel T. Larose, Chantal D. Larose, “Data Mining and Predictive Analytics”, 2015.
11. Venkat Reddy Korupally et al, “Big data analytics for Diagnosis and Prognosis of Cancer using Genetic Algorithm”, International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1251-1253.
12. Alamelumangai.N, Devishree.M, “An Ultrasound Image Preprocessing System Using Memetic ANFIS Method”, International Conference on Biology, Environment and Chemistry IPCBEE vol.1 (2011)IACSIT Press, Singapore.
13. Alamelumangai.N, Devishree.M, “PSO Aided Neuro Fuzzy Inference System for Ultrasound Image Segmentation”, International Journal of Computer Applications, 14-5, 2010.