ISSN:0975-9646

Big data analytics for Diagnosis and Prognosis of Cancer using Genetic Algorithm

Venkat Reddy Korupally[#1],Subba Rao Pinnamaneni[*2]

[#]*Assistant Professor, Dept of CSE, DRK Institute of Science & Technology*

**Associate Professor, Dept of CSE, Chalapathi Institute of Engineering & Technology*

Abstract - **Diagnosis and Prognosis are the two major challenging aspects which are to be addressed in treating cancer. The survival of Cancer patients depend upon the diagnosis of Cancer at the early stages (either in Stage I or Stage II). If the cancer diagnosed in Stage III or later stages, the chances of survival of the patient will become more critical. Prognosis will reveal the survival pattern for different attributes i.e., for specific drug, before and after the treatment. Better the diagnosis and prognosis, better the treatment outcome For Cancer. Generally single patient records will generate a large amount of data if we manage and analyze such big data, we may solve many problems in identifying the patterns which will lead to diagnose and prognosis of the cancer. This will help the doctors to take proper decisions. In this paper I am proposing a prototype which will be implemented on Hadoop and I am proposing a genetic algorithm which will analyze the open Cancer patient's data given by The Cancer Genome Atlas (TCGA) Data Portal, The National Cancer Institute, USA and guide the doctors in decision making in Diagnosis and Prognosis of the Cancer Patients.**

Keywords— **Big data, Hadoop, Genetic Algorithm**

I. INTRODUCTION

Amount of data generated every day is expanding in drastic manner. Big data is a popular term used to describe the data which is in zetta bytes. Government, companies many organizations try to acquire and store data about their citizens and customers in order to know them better and predict the customer behaviour. Social networking websites generate new data every second and handling such a data is one of the major challenges companies are facing. Data which is stored in data warehouses is causing disruption because it is in a raw format, proper analysis and processing is to be done in order to produce usable information out of it.

To analyze this much of huge data, we have suitable technologies include 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. Additional technologies being applied to big data include 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. Even though there are many suitable technologies to solve Big Data challenges, I want to explore a handful of the more popular ones, in particular the Amazon Cloud, Hadoop's MapReduce, and other open-source parsers.

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.

- Design a Prototype using HDFS which will effectively collect, clean, analyze the data from different data sources
- Developing an algorithm which can produce accurate results for the better Decision making in Diagnosis and Prognosis of the Cancer Patients
- The main objective of this project is to help the doctors in saving the lives of millions of people who are suffering from different types of Cancers worldwide.

III. GENETIC ALGORITHM

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomised, GAs are by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution, specially those follow the principles first laid down by Charles Darwin of "survival of the fittest.".

IV. IMPLEMENTATION DETAILS OF GENETIC ALGORITHM

Based on Natural Selection

After an initial population is randomly generated, the algorithm evolves the through three operators:

1.selection which equates to survival of the fittest;
2.**crossover** which represents mating between individuals;
3.mutation which introduces random modifinca -tions.

1. Selection Operator

i. key idea: give prefrence to better individuals, allowing them to pass on their genes to the next generation.
ii. The goodness of each individual depends on its fitness.
iii. Fitness may be determined by an objective function or by a subjective judgement.

2. Crossover Operator

i. Two individuals are chosen from the population using the selection operator

ii. A crossover site along the bit strings is randomly chosen

iii. The values of the two strings are exchanged up to this point

If S1=000000 and s2=111111 and the crossover point is 2 then S1'=110000 and s2'=001111

iv. The two new offspring created from this mating are put into the next generation of the population By recombining portions of good individuals, this process is likely to create even better individuals

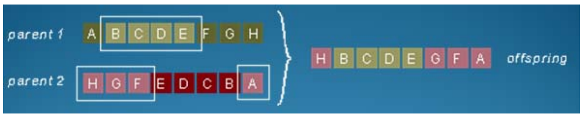

Fig i

3. Mutation Operator

i. With some low probability, a portion of the new individuals will have some of their bits flipped.

ii. Its purpose is to maintain diversity within the population and inhibit premature convergence.

iii. Mutation alone induces a random walk through the search space

Before Mutation

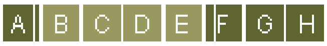

Fig ii

After Mutation

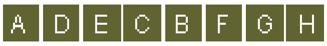

Fig iii

V. HDFS

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.

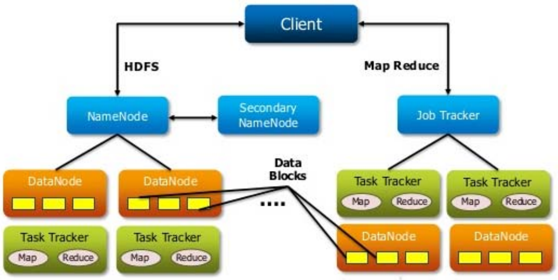

Fig iv

Assumptions and Goals

1) Hardware Failure

Hardware failure is the norm rather than the exception. An HDFS instance may consist of hundreds or thousands of server machines, each storing part of the file system's data. The fact that there are a huge number of components and that each component has a non-trivial probability of failure means that some component of HDFS is always non-functional. Therefore, detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

2) Large Data Sets

Applications that run on HDFS have large data sets. A typical file in HDFS is gigabytes to terabytes in size. Thus, HDFS is tuned to support large files. It should provide high aggregate data bandwidth and scale to hundreds of nodes in a single cluster. It should support tens of millions of files in a single instance.

3) Simple Coherency Model

HDFS applications need a write-once-read-many access model for files. A file once created, written, and closed need not be changed. This assumption simplifies data coherency issues and enables high throughput data access. A MapReduce application or a web crawler application fits perfectly with this model. There is a plan to support appending-writes to files in the future.

4) "Moving Computation is Cheaper than Moving Data"

A computation requested by an application is much more efficient if it is executed near the data it operates on. This is especially true when the size of the data set is huge. This minimizes network congestion and increases the overall throughput of the system. The assumption is that it is often better to migrate the computation closer to where the data is located rather than moving the data to where the application is running. HDFS provides interfaces for applications to move themselves closer to where the data is located.

VI. THE ALGORITHM

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 population(t) on Big data
ii. Determine fitness of population(t)
iii. Repeat
    a. select parents from population(t)
    b. perform crossover on parents creating population
    c. perform mutation of population
    d. determine fitness of population

until best individual is good enough
```

VII. CONCLUSION

Here in this paper i am analysing the Bigdata(Cancer Patients Data provieded by National Institute of Cancer) by applying the Genetic Algorithm Concepts on it. So that the doctors can come to clear ideas on the huge patients data and they can get some very good patterns to prognosis and diagnosis of the different types of Cancers. It is a very good area for the further research on Cancers so that combinely we may all reduces the cases of deaths casuing by different Cancers worldwide.

REFERENCES:

1. MA.Jabbar, B.L Deekshatulu, Priti Chandra, ”Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm”, International Conference on Computational Intelligence: Modeling Techniques and Applications
2. Patel, V. Adhil, M. ; Bhardwaj, T. ; Talukder, A.K. “Big data analytics of genomic and clinical data for Diagnosis and Prognosis of Cancer”, Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference
3. (CIMTA) 2013 "Home-PubMed-NCBI" , *United States National Library of Medicine (NLM)* [online] Available: www.ncbi.nlm.nih/ pubmed/
4. H. Fang and J. Gough, "The'dnet' approach promotes emerging research on cancer patient survival", *Genome Medicine 2014*, vol. 6, no. 64, 2014
5. L. Li, J. R. David, J. P. Chirag, C. W. Susan, C. Rong, P. T. Nicholas, T. D. Joel and J. B. Atul, "Disease Risk Factors Identified Through Shared Genetic Architecture and Electronic Media Records", *Sci Transl Med*, vol. 6, no. 234ra57, 2014
6. S. C. Pingle, Z. Sultana, S. Pastorino, P. Jiang, R. Mukthavaram, Y. Chao, I. S. Bharati, N. Nomura, M. Makale, T. Abbasi, S. Kapoor, A. Kumar, S. Usmani, A. Agrawal, S. Vali and S. Kesari, "In silicomodeling predicts drug sensitivity of patient-derived cancer cells", *Journal of Translational Medicine*, pp. 12-128, 2014
7. K. Goh, M. Cusick, D. Valle, B. Childs and M. Vidal, "The human disease network", *National Academy of Sciences*, vol. 104, no. 8685, 2007
8. W. Tian, L. Zhang, M. Taan, F. Gibbons and O. King, "Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function", *Genome Biology 9 Suppl*, vol. 1, no. S7, 2008
9. I. Ulitsky and R. Shamir, "Identification of functional modules using network topology and high throughput data", *BMC systems biology*, vol. 1, no. 8, 2007
10. N. Nagarajan, A. Tewari, J. O. Woods, I. S. Dhillon, E. M. Marcotte and U. M. Singh-Blom, "Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses", *PLOS one*, 2013