Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 7, July 2019, pg.79 – 87

COMPARATIVE STUDY OF GENETIC AND RANDOM FOREST ALGORITHM ON BONE MARROW GENE SEQUENCES

Dr. M.Mayilvaganan[1]; S.Sowmya[2]
Associate Professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India[1]
Email Id – mayil24_02@yahoo.co.in
Research Scholar, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India[2]
Email Id – sowmyasubramaniam.cbe@gmail.com

ABSTRACT: ***Bone marrow cancer is formed inside the spongy tissue found in the centre of bone. In Humans bone marrow are located in the ribs, vertebrae, sternum, and bones of the pelvis it contains stem cells that form into numerous types of blood cells found in the body namely red blood cells, white blood cells and platelets. When these cells grows too fast or abnormally results in Bone marrow cancer. DNA sequencing is the process of finding the accurate order of nucleotides in chromosomes and genomes. The development of techniques to store and search DNA sequences has led to widely applied methods especially string matching algorithms, machine learning and database theory.***

In this research work gene sequencing of bone marrow cancer is used to analyze and evaluate the performance of data mining techniques. The proposed work focused on comparative study of data mining algorithms. The Random forest and genetic algorithm were used as the evaluation indicators for the comparative study of the execution time and memory efficiency of each algorithm. The performance is analyzed based on the different number of instances and confidence in gene sequence data set.

Keywords: Data Mining, Machine Learning, DNA sequence, Bone marrow Cancer cell, Genetic and Random Forest algorithm.

I. INTRODUCTION:

Data mining refers to mining the knowledge from a large database. Data mining is most popularly used as knowledge discovery from data in Knowledge Discovery Data (KDD). The other view of data mining is one of the main steps in process of knowledge discovery. Data mining appliance will play out the information investigation and may reveal vital information designs, contributing extraordinary to business systems, learning bases, and logical and medicinal research. The enlarge gap between data and information is utilized for precise

advancement of data mining instruments that will transform the information tombs into "Brilliant pieces" of learning.

Data mining is the technique extracting data for utilization of taking in examples and models from expansive broad datasets. Data mining it include the employments of machine learning, measurements, artificial intelligence, database sets, design acknowledgement and perception. The Data mining process is not simply constrained to bioinformatics and is utilized in many contrasting businesses to give data intelligence. Bioinformatics deals with storage, gathering, simulation and analysis of biological data involving the sequences, molecules, gene expressions and pathways.

II. RELATED WORK:

- Author Bhavani K described to compare the efficiency of two data mining algorithms in Human Liver DNA data sequences. First Genetic algorithm using selection method and Next Decision Tree algorithm and often results about the performance, speed and memory accuracy [2].
- Author Mayilvaganan M, Hemalatha R attempt to analyze the DNA gene cancer dataset with RBC, WBC and Platelets cancer data set. Used association and classification rule based on FSA red algorithm with bloom filters and Apriori algorithm using hierarchical clustering are compared using data mining technique. Comparisons of algorithms are made on the execution time and memory efficiency [5].
- Author Amie Judith Radenbaugh attempt to compare the DNA and protein sequences with Genetic algorithm and Dynamic programming. Resulted in Genetic algorithm is better than DP in time of execution, performance increased, memory location was decreased and the implementation reduced the time [7].
- Author Ruchi Gupta, Dr.Pankaj Agarwal, Dr.A.K.Soni discussed multiple sequence alignment is carried out by developed genetic algorithm with new evaluation process, good mutation probability, and new recombination operators. Execution time that causes the stop condition so active compared to other algorithm techniques [8].
- Author Mayilvaganan M, Rajamani R discuss about the rule based classifier implemented to analysis the sequence between normal liver cell and cancer cell using First Apriori Algorithm to discover patterns with frequency above the minimum support threshold. The algorithm calculates rules that express probabilistic relationship between items in frequent item sets. Next FSA red algorithm performed a few of reduction techniques such as

attribute selection, row selection and feature selection. Often the result cells position occurrence, performance, speed and memory accuracy and distance matrix ratio analysis [10].

III. PROPOSED METHODOLOGY:

The methodology is using classification and prediction techniques of Machine Learning. A sample of gene sequence obtained from NCBI website and available to the UCI is used to carrying out the experiment. Machine learning is an information finding methods that instructs computer to do what works out easily to gain for a fact. The Classification and regression technique are used to diagnosis the computational biology for the detection and helps to take better decision and predictions.

3.1 Genetic Algorithm:

The genetic algorithm is a process for finding both constrained and unconstrained problems that is based natural selection that operates biological evolution. The genetic algorithm continuously changes a population of individual solution. At each step, the genetic algorithm chooses individuals at random from current population to be parents and use them to produce the children for next generation.

The genetic algorithm uses the following set of rules at each step to develop the next generation from current population:

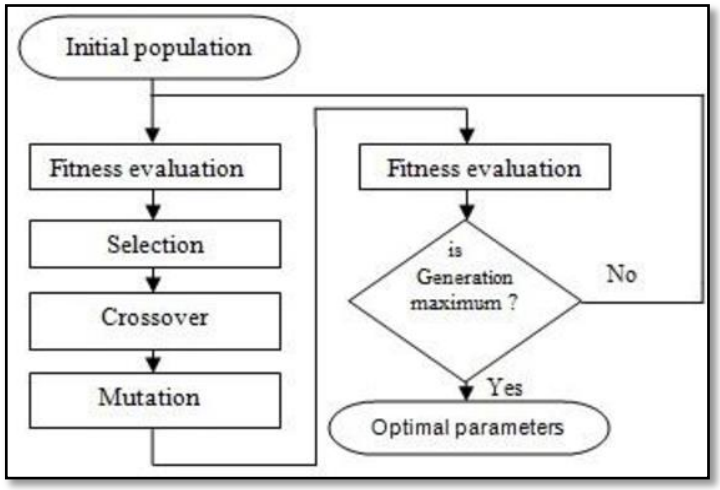

Figure 3.1 Flow of Genetic Algorithm operators

- Initialization: Define the population that contains individuals, each having their own set of chromosome.
- Fitness function: Defined function which takes candidate solution to the problem as input and produces as output.

$$F(x) = g(f(x))$$

- Selection Rule: Select the individuals, known as parents that shared to the population at the next generation.
- Crossover Rule: Combine two parents to form children for the next generation.
- Mutation Rule: Apply random changes to individual parents to form children.

3.2 Random Forest Algorithm:

Random forest is a machine learning method for classification, regression and another task that perform the construction of multiple decision trees at training time and output the class that is the classification of the classes or regression of the individual trees.

Random forest combines the "bagging" concept and arbitrary selection of features and independently to construct a collection of decision trees with controlled variance. The method of building a forest unrelated trees using a CART, combined with randomized node optimization and bagging. The basis of the modern practice of random forest is as follows:

- Using out-of-bag error to estimate the generalization error.
- Calculating variable concern through permutation.

Random Forest is the gathering of CART where every choice tree is completely developed till the terminal hub and the forecast from each tree is figured and the normal of the expectation of individual tree is ascertained to frame the woodland. The dataset are utilized for testing reason likewise got out of pack which are utilized to evaluate the Out of Bag (OOB) mistake for grouping.

The quantity of mtry highlights chose indiscriminately is constantly steady in the improvement of the tree and the timberland. Random Forest is the gathering of trees yet every one of the trees are completely developed without pruning.

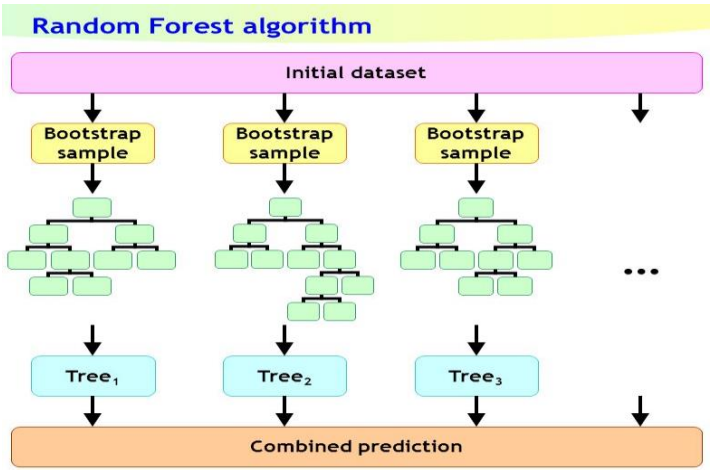

Figure 3.2 Flow of Random Forest Algorithm

A. Data for Research

Bone Marrow cancer - Homo sapiens C2 calcium dependent domain containing 6 (C2CD6), transcript variant 1, mRNA

NCBI Reference Sequence: NM_001168221.1

Definition- Homo sapiens C2 calcium dependent domain containing 6 (C2CD6),transcript variant 1, mRNA.

Accession - NM_001168221

Version- NM_001168221.1

Keywords source- Homo sapiens (human)

ATGGAGCCACCCCAAGAGACAAATAGGCCTTTCAGCACACTTGATAACCGCAGCGGGCAGGTCCA
AGTCCTGTCCGCCACCCCGCTATTGCAGAGGAATCCCTACAGCAGCCCGGATATCATGCACATTAA
AGGGTCGGAGGCTTCTTCGGTCCCTTACGCCCTGAACCAGGGCACGACGGCCCTGCCTAAGAACA
AGAACCAGGAGGGCACCGGGCACCGGCTGCTGAACATGCTGAGAAAAACTCTTAAAGAGTCTGAT
AGTGAAGAACTGGAAATAACACAGGAGACACCAAATTTGGTGCCATTTGGGGATGTGGTTGGCTG
CCTGGGCATTCATATAAAGAACTGCAGACATTTTATGCCTAAGATCAGTTTACAACATTATGCTAA
TTTATTCATTCGCATCTCTATAAACAAAGCTGTGAAATGTACAAAAATGTGTAGCTTGCTATCCAA
AAACGATGAGAAGAACACTGTAATTAAGTTCGATGAAGTGAAGTATTTTTCTGTACAGGTTCCCAG
ACGTTATGATGATAAGCGGAATAATATTTTATTGGAACTCATACAATATGACAATAGAGAAAAAC
GTGCTTTCTTGTTAGGGAGTGTTCAGATACATCTTTATGAGGTAATTCAGAAAGGGTGCTTCATTG
AAGAGGTCCAAGTGTTGCATGGAAACATATTTGTCTGCAGGCTGGAGGTGGAATTTATGTTCTCCT
ATGGAAACTTTGGTTATGGATTTTCACATCAGTTAAAACCTCTTCAGAAAATTACTGAGCCATCCA
TGTTTATGAATCTTGCACCACCTCCAGAAAGAACAGATCCCGTGACAAAAGTTATTACACCACAGA

CAGTAGAATATCCAGCATTCTTATCCCCAGACCTGAATGTTACTGTTGGGACTCCAGCTGTGCAAT
CCTCCAACCAGCCATCTGTAGTGCGACTTGAAAAACTTCAGCAACAACCCCGGGAAAGGCTTTAA
AGGGTCGGAGGCTTCTTCGGTCCCTTACGCCCTGAACCAGGGCACGACGGCCCTGCCTAAGAACA
AGAACCAGGAGGGCACCGGGCACCGGCTGCTGAACATGCTGAGAAAAACTCTTAAAGAGTCTGAT
AGTGAAGAACTGGAAATAACACAGGAGACACCAAATTTGGTGCCATTTGGGGATGTGGTTGGCTG
CCTGGGCATTCATATAAAGAACTGCAGACATTTTATGCCTAAGATCAGTTTACAACATTATGCTAA
TTTATTCATTCGCATCTCTATAAACAAAGCTGTGAAATGTACAAAAATGTGTAGCTTGCTATCCAA
AAACGATGAGAAGAACACTGTAATTAAGTTCGATGAAGTGAAGTATTTTTCTGTACAGGTTCCCAG
ACGTTATGATGATAAGCGGAATAATATTTTATTGGAACTCATACAATATGACAATAGAGAAAAAC
GTGCTTTCTTGTTAGGGAGTGTTCAGATACATCTTTATGATTAAAGGGTCGGAGGCTTCTTCGGTCC
CTTACGCCCTGAACCAGGGCACGACGGCCCTGCCTAAGAACAAGAACCAGGAGGGCACCGGGCAC
CGGCTGCTGAACATGCTGAGAAAAACTCTTAAAGAGTCTGATAGTGAAGAACTGGAAATAACACA
GGAGACACCAAATTTGGTGCCATTTGGGGATGTGGTTGGCTGCCTGGGCATTCATATAAAGAACTG
CAGACATTTTATGCCTAAGATCAGTTTACAACATTATGCTAATTTATTCATTCGCATCTCTATAAAC
AAAGCTGTGAAATGTACAAAAATGTGTAGCTTGCTATCCAAAAACGATGAGAAGAACACTGTAAT
TAAGTTCGATGAAGTGAAGTATTTTTCTGTACAGGTTCCCAGACGTTATGATGATAAGCGGAATAA
TATTTTATTGGAACTCATACAATATGACAATAGAGAAAAACGTGCTTTCTTGTTAGGGAGTGTTCA
GATACATCTTTATGAAAAAAA

IV. RESULTS AND DISCUSSIONS:

As a result of comparison made between Genetic and Random Forest algorithms on gene sequences it have been found that Random Forest is more effective and efficient to yield a better result when comparing to Genetic algorithm.

Random Forest algorithm occupies less memory space and less execution time then Genetic algorithm. The Genetic algorithm uses more execution time and huge memory range to run and for storage.

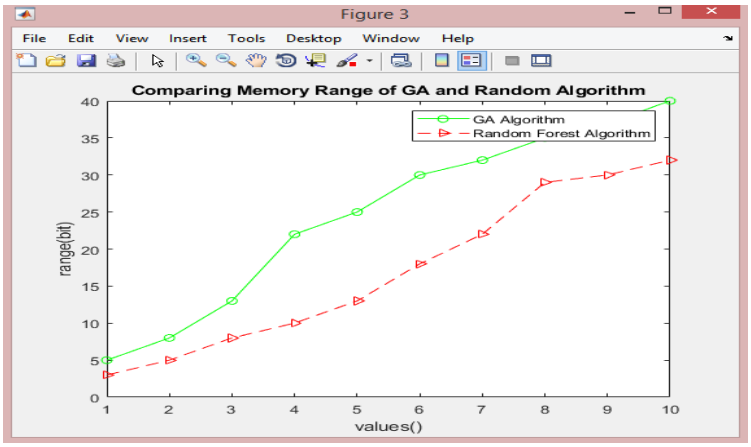

Figure 4.1 Comparing Memory Range of Genetic and Random forest algorithm

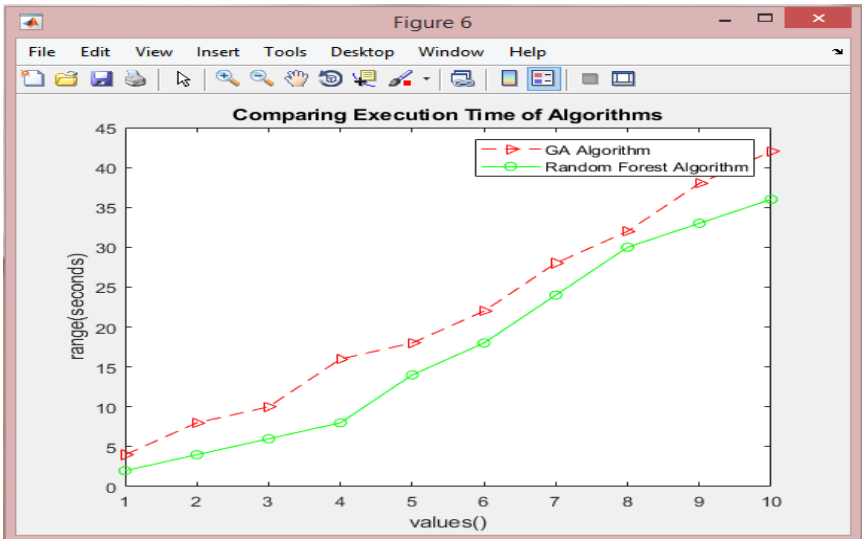

Figure 4.2 Comparing Execution Time of Genetic and Random forest algorithm

The following table present the result based on comparison of Genetic and Random Forest algorithm.

No.of.Datasets	Classifier Accuracy	Random Forest	Genetic Algorithm
5700	Time (Seconds)	24.57	26.95
5700	Memory Efficiency (GB)	505	581

V.CONCLUSION AND SCOPE OF FUTURE WORK

The data mining has a lot of functions to which can be used in order to analyze the gene sequences in large Dataset expressions. In this research work the goal was to present the usage of classification and prediction techniques to increase the possibilities of analysis in biological field. The main objective of this work was to evaluate the mining techniques to identify the one that will have the better performance in background of bioinformatics.

Based on the analysis performed on the data set, the following suggestions are planned for the future work:

Genetic algorithm can be combined with many different modifications such as selection method like Rank selection, Boltzmann selection, Random selection.

References

[1] Jiawei Han, Micheline Kamber "Data Mining Concepts and Techniques" in proceeding of second edition Morgan Kaufmann Publisher An imprint of Elsevier 2006.

[2] Bhavani K "Comparative Study of Genetic and Decision Tree algorithms on Gene Sequences".

[3] Dr Megan Y Murray "Investigating Cancer Stem Cells In Multiple Myeloma" "https://www.humaneresearch.org.uk/research/bone-marrow-cancer-research".

[4] Saurabh Sindhu, Divya Sindhu "Data Mining and Gene Expression Analysis in Bioinformatics" (IJCSMC) International Journal of Computer Science and Mobile Computing, Vol. 6, Issue.5, May 2017.

[5] Mayilvaganan M, Hemalatha R "Performance Comparison of FSA Red & Apriori Algorithm's in Mutation Analysis" (IJCTT) International Journal of Computer Trends and Technology, Vol.17, November 2014.

[6] Mohammed Zakariah "Classification of genome data using Random Forest Algorithm: Review" (IJCTA) International Journal of Computer Technology & Applications, Vol.5. ISSN: 2229-6093.

[7] Amie Judith Radenbaugh "Applications of Genetic Algorithms in Bioinformatics" "http://scholarworks.sjsu.edu/etd_theses".

[8] Ruchi Gupta, Dr.Pankaj Agarwal, Dr.A.K.Soni "Genetic Algorithm Based Approach for Obtaining Alignment of Multiple Sequence" (IJACSA) International Journal of Advanced Computer Science, Vol.3, 2012.

[9] Huiqing Liu, Limsoon Wong "Data Mining Tools for Biological Sequences" Journal of Bioinformatics and Computational Biology, June 2003.

[10] Mayilvaganan M, Rajamani R "Role of Data Mining in Nucleotide Sequence of Normal and Cancer Affected Liver Cells" (IJRCAR) International Journal of Research in Computer Applications and Robotics, Vol.2, 2014 ISSN: 2320-7345.

[11] Mayilvaganan M, Rajamani R "Rule Based Classifier Analysis with Nucleotide sequence in Normal Liver Cells and Affected Liver Cells" (IJIRCCE) International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, June 2014.

[12] Andrew Chipperfield, Peter Fleming, Hartmut Pohlheim, Carlos Fonseca "Genetic Algorithm TOOLBOX for Use with Matlab".

[13] Meraj Nabi, Abdul Wahid, Pradeep Kumar "Performance Analysis of Classification Algorithms in Predicting Diabetes" (IJARCS) "International Journal of Advanced Research in Computer Science", Vol.8, April 2017 ISSN No: 0976-5697.

[14] Dr.Kunwar Singh Vaisla, Jagmohan Rana "Bioinformatics Tools & Application" by Uttarakhand State Biotechnology Department, Government of Uttarakhand, 2012.

[15] Ramon Diaz-Uriarte, Sara Alvarez de Andres "Gene Selection and Classification of microarray data using Random Forest" in proceeding of BioMed Central The Open Access Publisher Vol.7, 2006.

