Diagnosis of Hepatitis Virus using Decision Tree and Genetic Algorithm

Mariya Khatoon
M.Tech Student
Deptt of Computer Science & Engineering
K.N.I.T Sultanpur, U.P, India

Abhay Kumar Agarwal
Assistance Professor
Deptt of Computer Science & Engineering
K.N.I.T Sultanpur, U.P, India

ABSTRACT
The aim of data mining is to extricate valuable information from large database. Data mining is the trending field of computer science which provides the efficient techniques and algorithms, performs in different domain like medical, education, business organization, banking sector, bioinformatics etc. in this paper we will provide the better algorithms to predict the presence and absence of hepatitis virus. Comparative analysis is done in this paper of decision tree and genetic algorithm to diagnose the hepatitis virus. Decision tree algorithm is used to predict the possible decisions to diagnose the hepatitis virus. And genetic algorithm is used to optimize the result.

Keywords
Data mining, Decision Tree, Genetic Algorithm, Hepatitis Virus

1. INTRODUCTION
Now a day's people suffering from many hazardous diseases and hepatitis one of them. Every year many people affected from this liver inflammation. So we will provide the efficient algorithms to predict the virus. Firstly we use decision tree, is the tree like structure and give the results in the form of decisions. Decision tree is to classify and predict the model. Its working based on the nodes means from top to bottom every node defines some attribute and child node is present in the last layer of tree provides the results and internal nodes known as decision node [1].

The second algorithm is genetic algorithm, based on theory of "Darwin" means the fittest species can easily adopt the changes around the environment and survive. We can say survival of the fittest population is easy as compared to unhealthy population. Genetic algorithm is plays important role regarding search problem, complex problem and optimization problem. It is most attractive technology in data mining field nowadays [2].

So with the help of those algorithms i want to provide that genetic algorithm gives optimized result as compared to decision tree algorithm. Genetic algorithm performs well due to its well regulating operators.

2. PREVIOUS WORK
Shiueng-Bien Yang et al. [3] describes that the decision tree learning is a technique that uses tree structure to solve system identification problem. It has been shown that the problem of constructing optimal decision trees that require fewest tests required to classify an unknown sample is an NP complete problem. This study develops a new decision tree based on the genetic algorithm. The genetic algorithm is proposed to determine the number of branches for each node in the new decision tree.

Tanmay Singha et al. [4] give the classification approach for mining of data. They have designed the model which predicts the large volume of data set and optimizes the hidden pattern and knowledge from that.

Maziyar Grami et al. [5] proposed the method of mining association rule using genetic algorithm. Results showed that genetic algorithm is more effective and efficient in discovering association rules in a transactional database. In this paper genetic algorithm gives better result as compared to BBO, FP growth and apriori algorithm.

Du Haizhou et al. [6] proposed that Decision tree is broadly used in predictive model and classification. The introduction of Generalized Decision Tree (GDT) realized the joining of classification rules and decision tree structure. This paper provides the way of construction of generalized decision tree by using DNA coding genetic algorithm.

Chafik Belghit et al. [7] proposed genetic algorithm to predict characterization parameters for petroleum fraction. They developed an equation to predict critical properties for hydrocarbons and petroleum fractions. The equation gave very good results, thus gratifying the main objective of this work which is to provide a simple calculation correlation that characterizes pure hydrocarbons as well as petroleum fractions using two input parameters that can be easily obtained; the normal boiling point and specific gravity.

3. PROPOSED METHODOLOGY
This proposed two algorithms for mining the same data set with reference to medical domain first is decision tree algorithm, is mainly a classification algorithm. Second is genetic algorithm which is used for the optimization of result. It performs with the help of bio inspired operators like crossover, mutation, fitness function, natural selection etc. [8]. Both give the separate result. Proposed model follows the different step which is discussed below:

- Firstly collect data set from UCI machine learning repository and then divide the data in training dataset and testing data set using fit function.

- After that apply decision tree and evaluate the performance of decision tree and then apply genetic algorithm and evaluate the performance.

- Both algorithms give some output so compare the decision tree algorithm's result to genetic algorithms result.

- After the comparative analysis of the both algorithms, now we analyze that which algorithm gives better result and this is the part of result analysis.
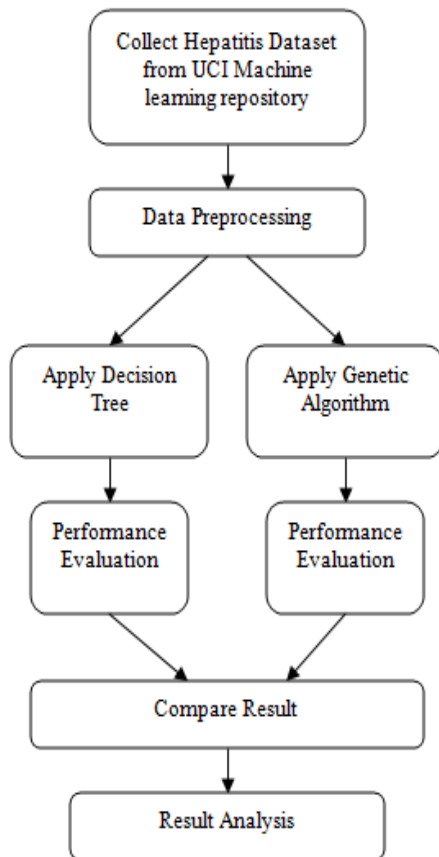
Fig.1. Proposed methodology

4. RESULT

4.1 Result using Decision tree

We have Develop a novel model for better result. Decision tree and genetic algorithm both perform well on hepatitis data set and give different results. There are 19 attributes which we have taken for experiment some are categorical and some are numeric attribute. There are two classes which show the absence and presence of hepatitis virus that live and die. Data set is collected from UCI Machine learning repository. Data preprocessing is done using fit function. After applying the decision tree classifier we got the possible decisions to diagnose the presence and absence of hepatitis virus, which is shown below-
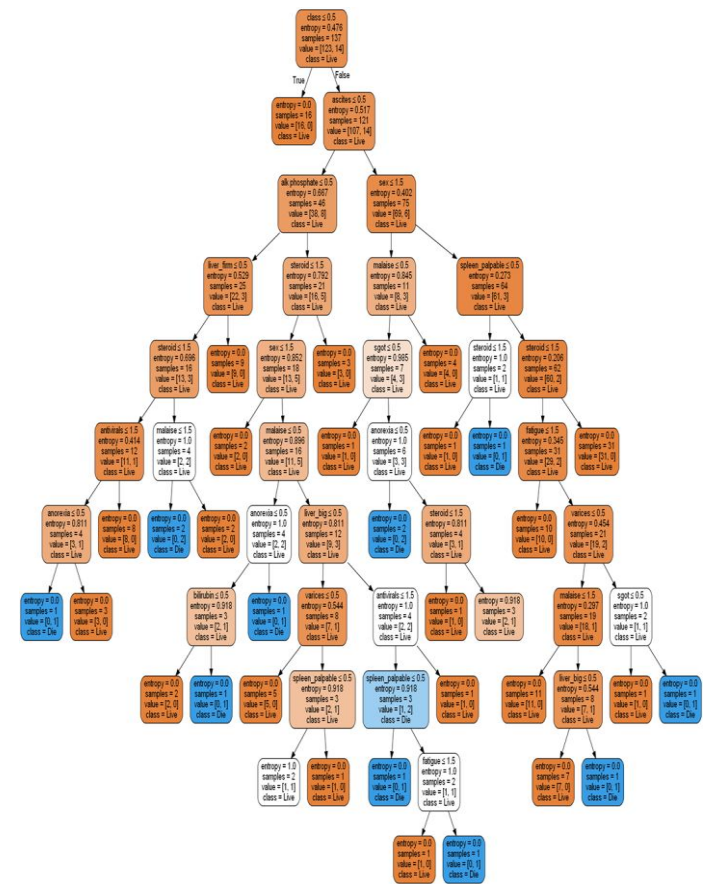
Fig.2. Possible decisions to diagnose hepatitis virus

4.2 Result using Genetic Algorithm

After the implementation of decision tree, implement genetic algorithm on same data set and get some results. In Genetic algorithm firstly initialize the population then calculate the fitness function to check the solutions quality and after that we apply crossover to produce the one or more offspring with the help of parent gene. And that mutation operation is done to maintain the genetic diversity form one population to another population. No of iterations continues in loop unless we didn't get the optimized result. Result using genetic algorithm is shown in table 1.

Table.1. Result using Genetic Algorithm

Current iterations	No of samples	Execution time in second	Accuracy in percentage
1	10	0.0	33.33
2	160	1.8	38.33
3	310	5.3	45.33
4	460	7.9	60.67
5	610	11.0	69.33
6	760	14.1	78.00
7	910	17.8	90.00
8	1060	21.5	94.00
9	1210	25.3	94.00

The above table shows the different results using genetic algorithm at 9th iteration we get optimized result.

5. COMPARATIVE ANALYSIS

In this paper we provide the comparative analysis of decision tree and genetic algorithm, with the help of this we shows that genetic algorithm perform well and give better and optimized result as compare to decision tree. Fig.3. shows the comparative study of both algorithms which is given below-
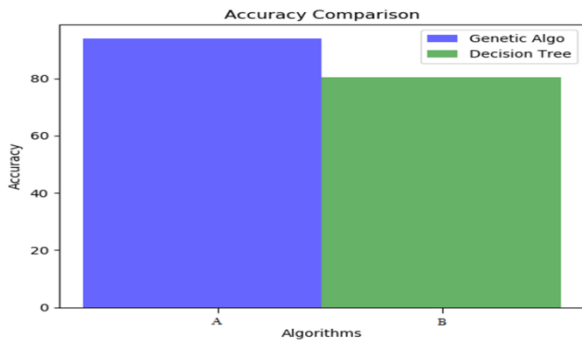
Fig.3. Comparative analysis of Decision Tree and Genetic Algorithm

The above Fig.3.provides that genetic algorithm give more accurate and optimized result. Diagnosis of hepatitis virus is liver inflammation and this paper become helpful to diagnose the virus easily. The whole experiment done on python technology and compatible tool used is pycharm. Table.2. shows the accuracy of both algorithms which is given below-

Table.2. Accuracy of Data Mining Algorithm

Applied Data Mining Algorithm	Recorded Accuracy in Percentage
Decision Tree	80.38
Genetic Algorithm	94.00

6. CONCLUSION AND FUTURE SCOPE

This paper proposed a model which is using two algorithms, decision tree and genetic algorithm. Both perform well on same dataset. But genetic algorithm gives better result comparatively due to its comprehensive and predictive rules.

The main purpose of this work is proposes the accomplished and faultless model for diagnosis of hepatitis virus. Such type of work helps in implementation and clinical support system.

Not only medical domain but also many other filed are present where data mining algorithms perform well. This work will provide the genuine result in diagnosis of hepatitis virus and becomes helpful to save many lives.

7. REFERENCES

[1] Mao guo-jun、Duan li-juan、Wang shi、Shi yun. "Principle and Algorithm of Data Mining" (Second Edition).Tsinghua University Press, 2007, pp120~121.

[2] G. Y. Yu, Y. Z. Wang, "Applied Research of improved genetic algorithms," Machinery, vol. 5, 2007, pp. 58-60.

[3] S. Yang and S. Yang, "New decision tree based on genetic algorithm," 2010 Int. Symp. Comput. Commun. Control Autom., vol. 1, pp. 115–118, 2010.

[4] T. Singha and S. Goswami, "Algorithm," 2017 Int. Conf. Energy, Commun. Data Anal. Soft Computer, pp. 3859–3864, 2017

[5] M. Grami, R. Gheibi, and F. Rahimi, "A Novel Association Rule Mining Using Genetic Algorithm," 2016 Eighth Int. Conf. Inf. Knowl. Technol., pp. 200–204, 2016.

[6] D. Haizhou, "Study on constructing generalized decision tree by using DNA coding genetic algorithm," 2009 Int. Conf. Web Inf. Syst. Min., pp. 163–167, 2009.

[7] C. Belghit, Y. Lahiouel, and T. A. Albahri, "Using A Real-coded Genetic Algorithm to Predict Characterization Parameters for Petroleum Fraction .," 2018 3rd Int. Conf. Pattern Anal. Intell. Syst., pp. 1–7, 2018.

[8] R. Kaur, "Multidimensional Attacks Classification Based on Genetic Algorithm and SVM," 2016 2nd Int. Conf. Next Gener. Comput. Technol., no. October, pp. 561–565, 2016.

[9] Zhi-Gen Hu, Jian-Ping Li, "Research And Application Of Data Warehouse And Data Mining Technology In Medical Field", 12th International Computer Conference On Wavelet Active Media Technology And Information Processing (ICCWAMTIP), IEEE 2015.