ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Genetic Algorithm for Classification of Data Mining Benchmark Dataset

[1] M. Kalai selvi, [2] Dr V. Lakshmi Praba
[1] M.Phil Scholar, Dept of Computer Science, Rani Anna Government College, Tirunelveli.
[2] Assistant Professor, Dept. of Computer Science, Rani Anna Government College, Tirunelveli.

Abstract: **- Data Mining or Knowledge Discovery is needed to make sense and use of data. The main goal of the data mining process is to extract data from a large datasets. Many algorithms are used for the solution of classification of data under the constraints. In this paper we use Genetic Algorithm (GA) for the classification of data. Genetic Algorithm finds the best fitness value that can find a feature set which yields classification rules. The process of reproduction and population replacement goes on until a stopping criterion is met. The paper demonstrates the strength and accuracy of this algorithm for White Wine dataset for classification in terms of performance efficiency and time complexity required.**

Keywords: **Accuracy, Classification, Data Mining, Fitness Function and Genetic Algorithm.**

I. INTRODUCTION

Knowledge Discovery is a non-trivial method of identifying valid, potentially novel helpful and ultimately understandable patterns in records. The Classification is one of the major role in Data mining. Data mining not only gathers and manages data but also includes analysis and prediction. Classification is a data mining method that assigns items in a collection to target categories or classes. The Genetic Algorithm consists of chromosomes that are made up of genes, which are individual elements that represent the problem. The collection of all chromosomes is called population. The Genetic Algorithm generates a population of points at each iteration. The algorithm repeatedly modifies the population of individual solutions. The optimal solution approaches from the best point in the population. The Genetic algorithm uses search strategy to find accurate and comprehensible knowledge within large database for the classification of dataset. The alteration and intersection of operators applied to the parents to produce their new off-springs. GA selects parents from the population set. These off-springs replaces the existing population set and the process is repeated to produce 'n' off-springs, where n is the size of population, and at the end of the iteration, the entire population is replaced by the new one. GA will evaluate each individual as a potential solution according to a predefined evaluation function. The suggested approach was tested with the White Wine dataset available from UCI machine learning repository in MATAB2014a for the classification of quality factors.

II. LITERATURE REVIEW

Surbhi Jain [1] proposed the Big Data using for clustering problems by combination of Genetic Algorithm. The Big Data concepts is useful for handling the large amount of datasets and different algorithm implementation. Genetic Algorithms are a people of computational prototypes stimulated by evolution theory of Darwin. According to Darwin the species which is fittest and can adapt to changing surroundings can survive; the remaining tends to die away. Darwin also affirmed that "the survival of an organism can be maintained through the process of reproduction, crossover and mutation. He also stated that the working mechanism is as follows: the algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population (reproduction). The reproductive prospects are distributed in such a way that those chromosomes which represent a better solution to the target problem are given more chances to reproduce than those which represent inferior solutions. Ranno Agarwal [2] explained the Genetic Algorithms are used in various fields of data mining technique to get the optimization solution for better performance and process the accurate result. Data mining is one of the important application fields of Genetic Algorithm. It provides a comprehensive search methodology for machine learning and optimization. Genetic Algorithm is in progress with a set of solutions called population. Solutions from one population are taken and used to generate a new population. This is provoked by a hope, that the new population will be better than the old one.

Pramod Vishwakarma, Yogesh Kumar, Rajiv Kumar Nath [3] stated in their research that, GAs is a search algorithm based on the natural selection and genetics. It uses a number of artificial individuals looking through a complex search space by using functions of selection, crossover and mutation. GA is use to finding optimal solution. This solution will conceal in a huge search space to look through. There is no guaranty to find any exact solutions when using a GA. In the proposed model the formal knowledge discovery in database process (KDD) is adopted to perform the data mining task, to get the interesting patterns or knowledge from the dataset. A. K. Santra, C. Josephine Christy [4] has proposed the clustering based on Niching Memetic algorithm and Genetic Algorithm to find the optimal solution and implement the feature selection. The confusion matrix contains the number of instances are to be correctly classified is the sum of diagonals in the matrix; all others are incorrectly classified accurately. In general, the fitness of a rule is assessed by its classification accuracy on a set of training examples.

Atul Kamble[5] in his Incremental Clustering in Data Mining using Genetic Algorithm. The research paper concentrated on new way of clustering using biological inspired Genetic Algorithm. This algorithm clusters data in dynamic form. The dataset is assumed to be clustered initially, and every new element is added as without any changing existing clustered database. A generation is over after each individual in the population has performed the genetic operators. The individuals of population will be better adapted to the fitness function, as they have to survive in the subsequent generations. At each and every step the GA randomly selects the individuals from the current population to produce the next generation. The generation completed successively, the population evolves toward an optimal solution.

Mamta Mor, poonam Gupta, Priyanka Sharma [6] presented in their study about the objective of fitness function is to maximize inter-cluster distance and minimize intra-cluster distance. The objects are clustered on the basis of Euclidean distance, each object belongs to the cluster whose centroid to object Euclidean distance is minimum. The GA design overcomes the two major drawbacks of k-means clustering algorithm i.e. converging at sub optimal result due to bad seed initialization. K. Sindhya, Dr. R. Rangaraj [7] proposed that the CKD based on Genetic Algorithm. This algorithm repetitively modifies a population of individual solutions. At each step, the genetic algorithm individually selects at random from the current population and uses them to produce the next generation. The Genetic Algorithm is used for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution.

M.Akhil jabbar, B. L Deekshatulu, priti chandra [8] proposed the combination of Genetic Algorithm and KNN to improve the classification accuracy of heart disease data set. The classifier is trained to classify heart disease data set as either healthy or sick. In this paper the approach could not account for irrelevant and redundant attributes present in above mentioned data sets. Cross over rate for GA have to be high and 60% is preferable ,so we set the value at 60%.Mutation rate have to be low and we set mutation value at 0.033.Population size have to be good to improve the performance of GA, so population size is fixed at 20. Keshavamurthy B. N, Asad Mohammed Khan & Durga Toshniwal [9] presented work to improves the rule based Genetic Algorithm classifier by improve upon the fitness function parameter modification. Crossover selects genes from parent chromosomes and creates a new offspring. Mutation changes randomly the new offspring. A fitness function can be defined by combining Confidence and Completeness. The key factor of GA is its fitness function, the convergence of search space is directly proposal to the effectiveness of fitness function in other words if fitness function is good then better the convergence of GA for a given problem. Pooja Goyal, Saroj[10] discovered the issues and challenges of applying GA based approaches for discovery of classification rules. It summarizes the manner in which the state-of-the-art research has addressed the issues like setting of GA parameters, seeding the population and speciation, local convergence and computationally expensive fitness computations etc. The paper also points to the research directions to enhance the efficacy and efficiency of GAs in the domain of classification rule discovery.

III. METHODOLOGY

A. Genetic Algorithm

The Genetic Algorithm is used to produce the new generations from the existing generation.

A Simple generational genetic algorithm procedure is given below.

- Choose the initial population of individuals
- Evaluate the fitness of each individual in that population
- Repeat on this generation until termination (time limit, sufficient fitness achieved)
- Select the best-fit individuals for reproduction
- Reproduce new individuals through crossover and mutation operations

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

- Assess the individual fitness of new individuals
- Replace least-fit population with new individuals

Initially process is started with a population value, in this experiment. The population value is set to have 50% accuracy for 100 generation, 40% accuracy for 50 generation and 50% accuracy for 30 generation.

The Parameters fixed for the algorithm are tabulated in Table 1.

Parameters	Value
Problem type	'bound constraints'
Rngstate	[1x1 struct]
Generations	51
Funccount	2700

IV. DATASET DESCRIPTION

The goal of White Wine dataset is to model Wine quality based on physicochemical tests. This dataset can be used for classifying the wine quality as 'Good' or 'Bad'.

Dataset characteristics : Multivariate
Attribute characteristics : Real
Associated Tasks : Classification
Number of instances : 4898
Number of Attributes : 12

Attribute Information:
Input variables

1. Fixed acidity ranges from 3.80 to 15.90
2. Volatile acidity ranges from 0.08 to 1.58
3. Critic acid ranges from 0.000 to 1.660
4. Residual sugar ranges from 0.60 to 65.80
5. Chlorides ranges from 0.009 to 0.611
6. Free sulfur dioxide ranges from 1.0 to 289.0
7. Total sulfur dioxide ranges from 6 to 440
8. Density ranges from 0.987 to 1.039
9. PH ranges from 2.72 to 4.01
10. Sulphates ranges from 0.220 to 2.000
11. Alchohol ranges from 8.0 to 14.9

Output variable

12. Quality ranges from 3.00 to 9.00

Sample Dataset

	FixAcid	VolAcid	CitAcid	ResSugar	Chlorides	FreeSO2	TotalSO2	Density	pH	Sulphates	Alcohol	Quality
2	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
3	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
4	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
5	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
6	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
7	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
8	6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6
9	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
10	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
11	8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	6
12	8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	5
13	8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	5
14	7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	5
15	6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	7
16	8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	5
17	6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	7
18	6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	6
19	6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
20	7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	6
21	6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22	0.5	9.5	5
22	6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
23	6.4	0.31	0.38	2.9	0.038	19	102	0.9912	3.17	0.35	11	7
24	6.8	0.26	0.42	1.7	0.049	41	122	0.993	3.47	0.48	10.5	8

In this paper, we use only two attributes for classifying wine quality.

V. EXPERIMENTAL RESULT

The main objective of Genetic Algorithm is used to find best fitness value. Proper mixing of contents (attributes specified) will produce the best quality Wine. In this work, initially only the first two attributes are considered for determining the best fitness value. Since the quality of Wine is determined by both the attributes, Wine is classified as 'Good' or 'Bad' based on both these attribute's best fitness value. Applying Genetic Algorithm for calculating best fitness function, the following result is obtained. Best fitness value for first attribute-fixed acidity is 7 and that second attribute-volatile acidity is 0.27. After applying GA, the dataset is classified as 'Good Wine' and 'Bad Wine' for the considered attributes.

The obtained result is tabulated in table 2

Wine category	Values
Good Wine	938
Bad Wine	3960

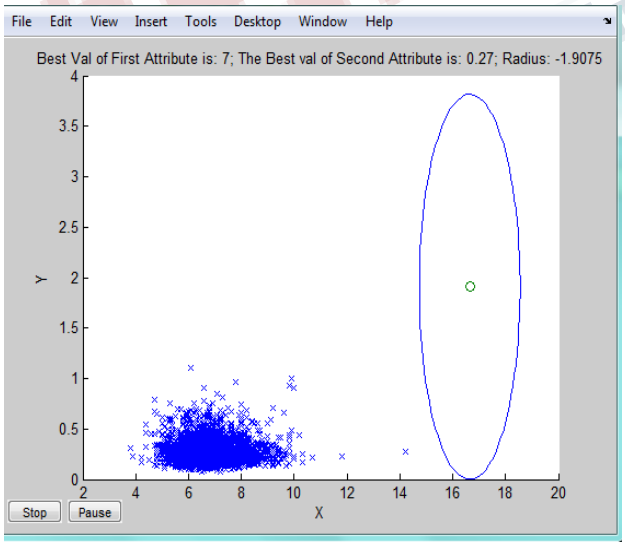
Screenshot for best fit value is in Figure 1

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

```
confusion matrix:
     5     0     0     0     0     0     0
    32     0     0     0     0     0     0
   291     0     0     0     0     0     0
   439     0     0     0     0     0     0
   176     0     0     0     0     0     0
    35     0     0     0     0     0     0
     1     0     0     0     0     0     0

accuracy = 50%
 The Best Value Predicted is
    7.0000    0.2700

------------Total No of good in Dataset-----------
   938

------------Total No of Bad in Dataset-----------
       3960

Elapsed time is 85.886209 seconds.
```

Graphical illustration of the obtained result using GA is in Figure 2



The obtained result is tabulated in Table 3

Resultant Type	Value
Predicted Output for the Population set	16.4692
Predicted Output for the Test Data set	3
Accuracy	50%
Elapsed time	85.886209
The Sum of Population	213.0177
The Best Fitness Value in the population	7.0118
Best Val of First Attribute	7
Best Val of Second Attribute	0.2700

VI. CONCLUSION

The aim of this paper is to classify the quality of Wine. In this process the classification of the White Wine dataset is done with two attributes namely fixed acidity and volatile acidity is based on the Genetic Algorithm. The Genetic Algorithm process is done on the White Wine datasets available from UCI machine learning repository in MATLAB2014a for the classification of quality factors. The obtained result from these attributes yielded 93.8% of Good quality Wine. In future all the attributes can be taken into consideration for determining the good quality wine.

REFERENCES

[1] Surbhi Jain," Mining Big Data using Genetic Algorithm", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 07 | July -2017.

[2] Ranno Agarwal, "Genetic Algorithm in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering", Vol.5, Issue 9, Sep-2015.

[3] Pramod Vishwakarma, Yogesh Kumar and Rajiv Kumar Nath," Data Mining Using Genetic Algorithm (DMUGA)", International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN : 2278-800X, Volume 5, Issue 2 (December 2012), PP. 88-94.

[4] A. K. Santra and C. Josephine Christy," Genetic Algorithm and Confusion Matrix for Document

Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012.

[5] Atul Kamble, "Incremental Clustering in Data Mining using Genetic Algorithm", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010 1793-8201.

[6] Mamta Mor1, Poonam Gupta and Priyanka Sharma," A Genetic Algorithm Approach for Clustering", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 6 June, 2014 Page No. 6442-6447.

[7] K. Sindhya and Dr. R. Rangaraj, "Design and Development of the Novel Genetic Algorithm Framework for Chronic Kidney Disorder Classification", International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307.

[8] M.Akhil jabbar, B.L Deekshatulu and Priti Chandra," Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, Procedia Technology 10 (2013) 85 – 94.

[9] Keshavamurthy B. N, A, Mohammed Khan and Durga Toshniwal," Improved Genetic Algorithm Based Classification", International Journal of Computer Science and Informatics (IJCSI) ISSN (PRINT): 2231 –5292, Volume-1, Issue-3.

[10] Pooja Goyal and Saroj, " Genetic Algorithms for Classification Rule Discovery: Issues and Challenges", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016.