Annals of Data Science (2023) 10(4):1007–1018
https://doi.org/10.1007/s40745-021-00354-9

Genetic Algorithms in the Fields of Artificial Intelligence and Data Sciences

Ayesha Sohail[1]

Received: 14 August 2020 / Revised: 5 June 2021 / Accepted: 30 July 2021 /
Published online: 30 August 2021

Abstract

In the fields of engineering and data sciences, the optimization problems arise on regular basis. With the progress in the field of scientific computing and research, the optimization is not a problem for small data sets and lower dimensional problems. The problem arise, when the data is large, stochastic in nature, and/or multidimensional. The basic optimization tools fail for such problems due to the complexity. The genetic algorithms, based on the natural selection hypothesis, play an imperative role to deal with such complex problems. Genetic algorithms are used in the literature to optimize numerous problems. In the field of computational biology, these algorithms have provided cost effective solutions to find optimal values for large data sets. The genetic algorithms have been used for image reconstruction. These algorithms are based on sub-algorithms to improve the accuracy and precision. We will discuss the advanced genetic algorithms and their applications in detail. Genetic algorithm, in hybrid form have attracted interest of researchers from almost all fields, including computer science, applied mathematics, engineering and computational biology. These tools help to analyze the systems in a swift manner. This important feature is discussed with the aid of examples. The time series forecasting and the Bayesian inference, in combination with the genetic algorithms, can prove to be powerful artificial intelligence tools. We will discuss this important aspect in detail with the aid of some examples.

Keywords Artificial neural network · Time series · Natural selection · Preprocessing · Randomness · Training algorithms

✉ Ayesha Sohail
ayeshasohail81@gmail.com

[1] Department of Mathematics, Comsats University Islamabad, Lahore Campus, Lahore 54000, Pakistan

1 Introduction

In the field of data science, with machine learning, several optimization tools have evolved over the last dacade [1–5]. In this manuscript, we will discuss the genetic algorithm, with a target in mind, to improve the previously reported studies, based on parametric optimization [6–11]. To understand the basis of the algorithm, that we will utilize during this research, we will first define the core concept in detail.

1.1 Genetics

Genetics is a branch of biology that deals with the genes and the genetic variations.

Gregor Mendel in the mid-19th century pioneered the work in the field of genetics, and is known as the "father of genetics". Before Mendel's period of revolution in the field of genetics, "genetics was primarily theoretical" where as afterwards, the science of genetics was evolved to "experimental genetics", resulting in tremendous progress in the field of bio-informatics.

A gene is defined as the basic unit of heredity, based on a sequence of "nucleotides in DNA or RNA". It is a section/part of the DNA made up of a sequence of *As*, *Cs*, *Ts* and *Gs*. There are around 20,000 types of genes in a human body.

1.1.1 Allele

An allele is defined as one of the multiple (two, or more) versions, of the "same gene" at the same place on a chromosome.

1.1.2 Gene Frequency

Let A_S be the number of a specified allele in a given population and A_T be the total of all alleles at its genetic locus. Then the gene frequency f is defined as:

$$f = \frac{A_S}{A_T} \tag{1}$$

1.1.3 Disturbance in Natural Equilibrium

In the absence of any disturbances, the gene frequencies usually remain constant from generation to generation.

The first three causes i.e. Mutation, migration, and the genetic drift are random processes. They may change the gene frequencies, without any regard to whether such changes increase or decrease the "likelihood of an organism's" survival or reproduction rate. Natural selection, on the other hand is different and is defined as:

1.2 Natural Selection

History has always admired the contributions of two pioneers in conceptualizing the natural selection process. These include "Alfred Russel Wallace" who was a naturalist who independently proposed the theory of evolution by natural selection, and was a great admirer of "Charles Darwin (naturalist)". Darwin and Wallace worked together as the co-proposers of the "evolution by natural selection".

The process of natural selection, can be better understood with the help of the quote: "The theory of natural selection revolutionised our understanding of living things, furnishing us with a comprehension of our existence where previously science had stood silent".

Natural selection helps in moderating the damage/randomness caused by the first three causes, since it has the ability to multiply the incidence of the "beneficial mutations" over the generations and eliminates harmful ones.

Thus the natural selection preserves the variations, that are most suitable in raising the chance of survival as well as the procreation. Such selections are multiplied from generation to generation, with least variations.

In other words, natural selection (survival of the fittest) is the reproduction and survival of individuals with "favorable traits". Thus the process of natural selection leads to evolutionary change.

1.3 Algorithm Based on Natural Selection

Genetic algorithms are search methods based on principles of natural selection and genetics. The genetic algorithms encode the decision variables of a search problem into finite length strings of alphabets of certain cardinality. The strings which are candidate solutions to the search problem are referred to as chromosomes, the alphabets are referred to as genes and the values of genes are called alleles.

For example, in an epidemiological problem, the area under consideration may be represented as a chromosome, whereas the sub regions of the area, based on the statistical features of the epidemic problem, will represent the gene.

1.4 Working Pattern of Genetic Algorithm

The genetic algorithms are mainly used to optimize the research problems, such as the feature selection, to model various aspects of the "natural immune system" and for other modern optimization problems, including machine learning and data science. A schematic is presented in Fig. 1. The basic optimization algorithms work in a different manner as compared to the genetic algorithms.

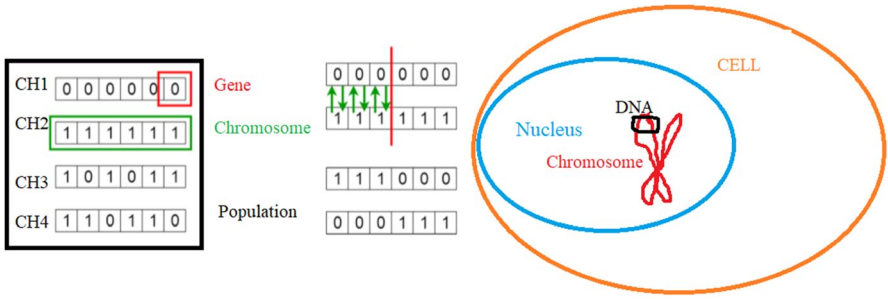

Fig. 1 Similarities in natural process (right) and algorithm strategy (left)

1.4.1 Basic Optimization Algorithms

Optimization is a term, extensively used in developing the "portfolios", trading, energy management, production planning and in almost all fields of sciences.

Optimization is a field of applied mathematics, where unconstrained and constrained problems are solved, to obtain the optimal solutions. The functions, matrices and variable expressions are used to define the underlying mathematics, of the given problem. Matlab, Pyhton, R or other programming softwares, are used to solve such optimization problems, using the automatic differentiation of the "objective" and "constraint functions" for swift evaluation of the optimal solution.

Optimal solutions can be obtained for the continuous as well as the discrete problems, to perform the "tradeoff analyses", and to incorporate the opti.mization methods into "algorithms and applications".

With the aid of good programming tools, one can design optimization tasks, including:

- Parameter estimation,
- Component selection,
- Parameter tuning.

1.4.2 Genetic Algorithms Versus Basic Optimization Algorithms

A genetic algorithm is defined as a method that is based on the natural selection process (as discussed in Sect. 1.2) and is used to solve the unconstrained and constrained optimization problems.

The algorithm is based on the concept of repeated modifications of the collection of individual solutions. At each step, the algorithm selects individuals randomly and use them to generate next iteration, until an optimal solution is obtained.

While comparing the basic optimization tools with the genetic algorithms, one observe that the basic tools generates a single point at each iteration and a

sequence of points approaches to the desired optimal solution. On the other hand the genetic algorithms generates a "population of points" at each iteration. At the end of the algorithm, the best point in the population converges to an "optimal solution".

Another important feature of the genetic algorithms is that they are stochastic in nature, whereas the standard optimization tools are deterministic in nature.

1.5 Application of Genetic Algorithm to Biological Problems

In the field of computational biology, the data based studies require serious attention and slight negligence or data generated errors may lead to false conclusions. The optimization algorithms, for the accurate data analysis have been used in the field of computational biology especially in the field of omics and radiomics.

In the next section, the mathematical basics of three different types of genetic algorithms are defined. Next, we have implemented the Matlab solvers on the biological datasets. The results are discussed in the next section and useful conclusions are drawn.

2 Materials and Methods

2.1 Materials

The genetic algorithm works in a step by step manner.

2.1.1 Step 1: Randomization

A random initial population is created during the first step.

2.1.2 Step 2: Reproduction

The algorithm then creates a sequence of new populations. In a recursive manner, current population is used to generate next population. To create the new population, the algorithm performs the steps as presented Fig. 2.

2.1.3 Step 3: Stopping Criteria

The purpose of genetic algorithms is to optimize the given problem, and thus the iterative procedure must achieve its finishing value, under the stopping criteria. A list of stopping criterion is provided in Fig. 3.

2.2 Methods

Genetic algorithms are used in the literature to optimize numerous problems. In the field of computational biology, these algorithms have provided cost effective

* Raw Fittness Score (**RFS**)
Scores each member of the current population by computing its fitness value.

* Scales **RFS** to convert them into Expectation Values (**EV**).

* Based on **EV**, selects members (parents).

*ELITE: are individuals in the current population with lower fitness.

* These ELITE individuals are passed to the next population.

* Produces children from the parents:

+ Either by making random changes to a single parent~~**mutation**
+ Or by combining the vector entries of a pair of parents~~**crossover**

* Next Generation (NG): by replacing the current population with the children.

Fig. 2 Sub-steps of step 2 of the genetic algorithm

MaxGenerations — The algorithm stops when the number of generations reaches MaxGenerations.

MaxTime — The algorithm stops after running for an amount of time in seconds equal to MaxTime.

FitnessLimit — The algorithm stops when the value of the fitness function for the best point in the current population is less than or equal to FitnessLimit.

MaxStallGenerations — The algorithm stops when the average relative change in the fitness function value over MaxStallGenerations is less than Function tolerance.

MaxStallTime — The algorithm stops if there is no improvement in the objective function during an interval of time in seconds equal to MaxStallTime.

FunctionTolerance — The algorithm runs until the average relative change in the fitness function value over MaxStallGenerations is less than Function tolerance.

ConstraintTolerance — The ConstraintTolerance is not used as stopping criterion. It is used to determine the feasibility with respect to nonlinear constraints. Also, max(sqrt(eps),ConstraintTolerance) determines feasibility with respect to linear constraints.

Fig. 3 A list of stopping criteria provided by Matlab^{TM}

solutions to find optimal values for large data sets. We will outline some recent features of genetic algorithms.

2.2.1 Bioinformatics: To Analyse Mass Spectrometry Data

In the field of bioinformatics, data sets include (1) “DNA sequences of genes”; (2) sequences of proteins (the proteins are usually 50 to 2000 amino acids long); (3)

protein tertiary structure data; (4) nucleic acids (size range varies from 21 nucleotides (small interfering RNA) to large chromosomes (human chromosome 1 is a single molecule that contains 247 million base pairs)) and (5) protein–nucleic acid complexes.

In addition to special skills and expertise, smart tool boxes are highly desired to explore the bioinformatics datasets.

Matlab community, "Mathworks", over the years has provided useful toolboxes, which are not only cost effective, but due to the builtin environments, are user friendly as well.

Genetic algorithms has been applied to explore the optimal features (peaks) in mass spectrometry data. To ditinguish between the cancer patients from the control patients, specific peaks were identified with the aid of genetic algorithm. "Biogacreate function" was used for this purpose. This function generates a "population matrix" with user defined dimensions. The details of this example can be found at (https://uk.mathworks.com/help/bioinfo/ug/genetic-algorithm-search-for-features-in-mass-spectrometry-data.html).

2.2.2 Hybrid Genetic Algorithm

The word hybrid refers to a combination of two or more algorithms to boost the efficiency of a solver. For the genetic algorithms to perform better, various hybrid algorithms have been used in the literature, see for example the review provided by El-Mihoub et'al. [12] and the references therein. Very recently, Drezner and Drezner [13] documented the significance of hybrid genetic algorithms in detail.

While using softwares, such as Matlab, R and Python, developing hybrid algorithms has made easy due to the emerging builtin toolboxes.

For example, Matlab provides a detailed example to build a hybrid algorithm for better optimal values of a given problem. This example bridges the genetic algorithm with the basic optimization tool. Since several functional evaluations are desired by the genetic algorithm, to approach the neighbourhood of the optimal value, it thus require time to perform effectively. One can speed up this procedure, by first generating a limited number of generations using genetic algorithm, and then utilize the resulting solution, as an initial guess for the basic optimization tool, to achieve optimal result, more rapidly.

The recommended hybrid toolbox is provided on (https://uk.mathworks.com/help/gads/when-to-use-hybrid-function.html).

2.3 Genetic Algorithms for Time Series Forecasting

Time series forecasting has remained an attractive field for the researchers from different disciplines such as computer science, computational biology and engineering. The forecasting algorithms helps in identifying the future of current research topics, based on the past and recent evidences.

Time series forecasting tools have been improved since the success of the neural network forecasting tools. The neural network tools are based on the input and output layers and on closed and open loops.

In the literature, genetic algorithms have been used to forecast the time series, such as the work provided by Khashei and Bijari [14], where they have verified with examples, the potential of genetic algorithm to become a powerful tool for "time series modelling and forecasting".

2.4 Genetic Algorithms for Bayesian Learning

The parameters in the Bayesian networks can be approximated with the aid of the genetic algorithms in association with the expectation maximization algorithm. An attempt has been made by the researchers [15], where global convergence of this algorithm was verified theoretically.

Feature selection has been used extensively, in the field of data mining, using different classification tools. Naive Bayes classifiers can be perform better with the genetic algorithms. Das et'al. [16] used set theory based algorithm for this purpose.

2.5 Bayesian Framework for Genetic Algorithm

Next, the genetic algorithms can perform better if a Bayesian framwork is selected. Since the genetic programming is iterative and consecutive population dependent, the Bayesian Gaussian process can make use of the "Bayes theorem" to find the "posterior distribution" (from the prior distribution) of programs. [17] presented two different genetic algorithms, derived from the Bayesian Gaussian process framework.

3 Results

3.1 Clustering via Genetic Algorithm for Colour Image Segmentation

Basically, machine learning tools are subdivided into two categories, the supervised and the unsupervised tools. Clustering tools are the unsupervised tools and are used for the natural grouping within a given data/image.

The genetic algorithms have been used in the recent studies for the image segmentation with the aid of evolutionary clustering. The objective function was based on the measured cluster distance and on the red, green and blue values (features).

Recently, the Matlab Mathworks toolbox has been equipped with such features, where the matrix formalism is used to develop and execute the algorithm.

The K-means clustering has been used extensively in the literature for the clinical data mining [6]. The K-means clustering basically while using the K-means clustering, the user picks "k" random data points/items from the given data and then label them as the "cluster-representatives". Next, the remaining data points are associated with the nearest cluster representatives. The euclidean distance/ norm is used

to calculate the similarities and differences among the clusters. The evolutionary genetic algorithms have much scope in this field since these are used to improve the performance of the method. The K-means principles for dividing objects into groups with high similarity, can be exploited, to propose a genetic algorithm. In such method, a population of chromosomes is evolved. Next, the one step K-means approach is used. Matlab Mathworks toolbox is equipped with a special function "gakmeans". This approach provides higher values of "evaluation indices" as compared to the K-means method, used in general.

Another important achievement of the genetic algorithms in the field of image segmentation is the binary image reconstruction. The binary image re-construction is a process, where the image is reconstructed with the aid of a small set of projections. The genetic algorithm uses the initial results from a linear back projection. This approach is used to reconstruct the images, to optimize the limiting values for the gray scale. The benefit of the genetic algorithm is that it avoids pixel by pixel optimization of the gray values. Thus the genetic algorithm based technique for the reconstruction converges quickly with a small number of iterations. Matlab Mathworks toolbox provides the users the following step by step routine for this purpose:

1. Crossover
2. Elitism
3. Fitness function
4. Genetic algorithm for optimization
5. Initialization
6. Mutation
7. Perpetration of the image and then selection.

3.1.1 Speedy Genetic Algorithm

Loops are used in coding and genetic algorithm development in all softwares, including Matlab, C/C++ and Java. Besides Matlab, other softwares are compatible with the nested loops. Matlab on the other hand, perform faster, if the loops are converted to array operations (vectorization). Thus Matlab indexing schemes can transform the codes with nested loops to easier, swift and short codes.

SpeedyGA was proposed in 2010 for fast and convergent application of genetic algorithm, using Matlab (https://uk.mathworks.com/matlabcentral/fileexchange/15164-speedyga-a-fast-simple-genetic-algorithm).

3.1.2 Artificial Neural Networks and Genetic Algorithms

Genetic algorithm is an optimization tool, that works with the random strings. These strings represent the design or the decision variables. Under the given constraints, the strings are analyzed for the fitness. The termination condition is used as the stopping criteria. If the termination criteria is not satisfied, the crossover, reproduction and the mutation strategy is adopted and a new population is created. This procedure continues till the desired output is obtained.

The genetic algorithm is used in the literature [18] to optimize the key players of the neural networks. These includes the bias, the learning rate and the activation constant.

In a recent study, the genetic algorithm was used to improve the neural network performance. For the verification of the improved performance, the results were also derived from the back propagation algorithm. It was concluded that the genetic algorithm worked better as compared to the back propagation, in improving the performance of the neural network.

3.1.3 Genetic Algorithms and the Convolutional Neural Networks

The Convolutional Neural Networks (CNNs) work as the image classification tool and their applications in different fields have received successful results. The basic building block of the convolutional neural networks is the "accurate architecture", and these architectures are sometimes manually designed. For manual working, rather than a computer based ready to use architecture, it is highly desired to have concrete knowledge of networking, mathematics and imaging. Such tasks can be tackled only by selected users. To faciliate the other users, who are not expert of this field, but wish to utilize the CNNs for image processing, a useful automatic tool, with the aid of genetic algorithm has been recently designed by the group led by Sun [19]. The "automatic characteristic" of their proposed approach helps the users, to obtain a promising architecture, for their specific problems, without worry about the domain knowledge of the CNNs. This approach outperformed over other architectures. The outcomes of the numerical experiments demonstrated that the proposed algorithm outperformed the existing automatic CNN architecture design algorithms in terms of (1) classification accuracy, (2) "parameter numbers" and (3) "consumed computational resources".

3.1.4 Genetic Algorithms for Modeling

In the field of modeling and simulations, the genetic algorithms and the hybrid genetic algorithms have always served to provide useful optimization benckmarks. Recently the work presented by Lin [20] showed the feasibility of the genetic algorithms for better approximation of the parameters involved in designing the A new optimization model of the "combined cooling, heating and power" CCHP system.

4 Discussion

The genetic algorithms are inspired from the nature's selection criteria of best survival. The genetic algorithms, if utilized properly, can help to explore a given problem more effectively, since its algorithm is based on three important of randomization, reproduction and stopping. The methods discussed in this manuscripts, outline the important applications of genetic algorithms in different disciplines. Different softwares freely provide toolboxes, for efficient implementation of genetic

algorithms. We have outlined some important and user friendly resources in this manuscript.

5 Conclusions

The future of genetic algorithms will see great success as their combined applications have already achieved milestones. In this manuscript, we have discussed the genetic algorithms in detail. We have explained the main motivation behind this algorithm and have discussed the importance of this approach in almost all data based studies and for evaluating the complex mutli-dimensional problems. We conclude that, the artificial intelligence tools such as the optimization tools, forecasting tools (such as time series) and the classification tools (such as the Naive Bayes classifiers), can perform better with the aid of genetic algorithms.

Authors' contribution AS did visulization, conception and modeling during this research.

Funding None.

Data and Code Availability All the data and material are provided within the manuscript. Code repositories are mentioned with in the manuscript.

Conflict of interest The authors declare that there is no conflict of interest.

References

1. AbuJarad MH, Khan AA, Khaleel MA, AbuJarad ES, AbuJarad AH, Oguntunde PE (2020) Bayesian reliability analysis of Marshall and Olkin model. Ann Data Sci 7(3):461–489
2. Gramaje A, Thabtah F, Abdelhamid N, Ray SK (2019) Patient discharge classification using machine learning techniques. Ann Data Sci 1–13
3. Olson DL, Shi Y, Shi Y (2007) Introduction to business data mining, vol 10. McGraw-Hill/Irwin, New York
4. Shi Y, Tian Y, Kou G, Peng Y, Li J (2011) Optimization based data mining: theory and applications. Springer, Berlin
5. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4(2):149–178
6. Sohail A (2019) Inference of biomedical data sets using Bayesian machine learning. Biomed Eng Appl Basis Commun 31:1950030
7. Iftikhar M, Sohail A, Ahmad N (2019) Deterministic and stochastic analysis of dengue spread model. Biomed Eng (Singapore) 31:1950008
8. Sohail A, Idrees M, Sajjad M, Iftikhar S, Tunc S (2020) Computational framework to explore impact of environmental stress on epidemics. Biomed Eng (Singapore) 32:2050047
9. Yu Z, Sohail A, Nutini A, Arif R (2020) Delayed modeling approach to forecast the periodic behaviour of sars-2. Front Mol Biosci 7:386
10. Yu Z, Ellahi R, Nutini A, Sohail A, Sait SM (2021) Modeling and simulations of CoViD-19 molecular mechanism induced by cytokines storm during SARS-CoV2 infection. J Mol Liq 327:114863
11. Yu Z, Arif R, Fahmy MA, Sohail A (2021) Self organizing maps for the parametric analysis of COVID-19 SEIRS delayed model. Chaos Solitons Fractals 150:111202
12. El-Mihoub TA, Hopgood AA, Nolle L, Battersby A (2006) Hybrid genetic algorithms: a review. Eng Lett 13(2):124

13. Drezner Z, Drezner TD (2020) Biologically inspired parent selection in genetic algorithms. Ann Oper Res 287(1):161
14. Khashei M, Bijari M (2011) A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl Soft Comput 11(2):2664
15. Sundararajan PK, Mengshoel OJ (2016) A genetic algorithm for learning parameters in Bayesian networks using expectation maximization. In: Conference on probabilistic graphical models, vol 1, p 511. PMLR
16. Das AK, Sengupta S, Bhattacharyya S (2018) A group incremental feature selection for classification using rough set theory based genetic algorithm. Appl Soft Comput 65:400
17. Zhang BT (2000) Bayesian methods for efficient genetic programming. Genet Program Evolvable Mach 1(3):217
18. Chen N, Xiong C, Du W, Wang C, Lin X, Chen Z (2019) An improved genetic algorithm coupling a back-propagation neural network model (IGA-BPNN) for water-level predictions. Water 11(9):1795
19. Sun Y, Xue B, Zhang M, Yen GG, Lv J (2020) Automatically designing CNN architectures using the genetic algorithm for image classification. IEEE Trans Cybern 50(9):3854
20. Lin H, Yang C, Xu X (2020) A new optimization model of CCHP system based on genetic algorithm. Sustain Cities Soc 52(101811):101811