Introduction to Genetic Algorithm and its Applications

Sathisha.G [1], Aaryan[2], Sheethal P.R[3], Khatija Faisal[4]
Department of Computer Science Engineering
[1,2,3,4] Atria Institute of Technology Bangalore, India

Abstract- *The Genetic Algorithm (GA) is a search heuristic that is routinely used to generate useful solutions to optimization and search problems. It generates solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Genetic algorithms are one of the best ways to solve a problem for which little is known. They are a very general algorithm and so work well in any search space.*
This article introduces the genetic algorithm (GA) as an emerging optimization algorithm. After a discussion of traditional optimization techniques, it reviews the fundamental operations of a simple GA and a number of applications, such as study of GA in sentiment analysis and GA based feature selection on Pima Indians diabetes that are being successfully implemented aredescribed.

Keywords- Diabetes Mellitus, Genetic algorithm, feature selection, Sentiment analysis tasks and framework.

I. INTRODUCTION

Genetic algorithms are search and optimization algorithms based on the principles of natural evolution, which were first introduced by john Holland in 1970. Genetic algorithms also implement the optimization strategies by simulating evolution of species through natural selections. Genetic algorithm is generally composed of two processes. First process is selection of individual for the production of next generation and second process is manipulation of the selected individual to form the next generation by crossover and mutation techniques. The selection mechanism determines which individual are chosen for reproduction and how many offspring each selected individual produce. The main principle of selection strategy is the better is an individual; the higher is its chance of being parent.

II. GENETICALGORITHM

Genetic algorithms (GA) are search algorithms based on the principles of natural selection and genetics, introduced by J Holland in the 1970's and inspired by the biological evolution of living beings. Genetic algorithms abstract the problem space as a population of individuals, and try to explore the fittest individual by producing generations iteratively. GA evolves a population of initial individuals to a population of high quality individuals, where each individual represents a solution of the problem to be solved. The quality of each rule is measured by a fitness function as the quantitative representation of each rule's adaptation to a certain environment.

The procedure starts from an initial population of .randomly generated individuals. During each generation, three basic genetic operators are sequentially applied to each individual with certain probabilities, i.e. selection, crossover and mutation.

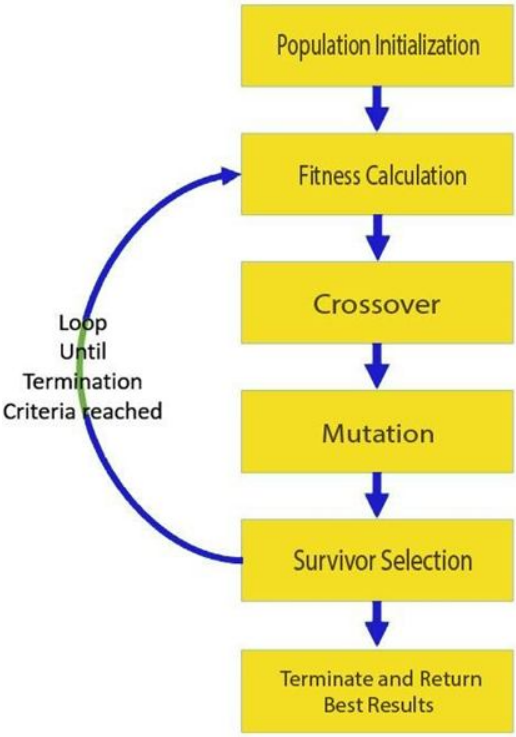
Population Initialization
Fitness Calculation
Crossover
Mutation
Survivor Selection
Terminate and Return Best Results
Loop Until Termination Criteria reached

Figure 1. Flowchart of GA System

The GAs is computer program that simulate the heredity and evolution of living organisms. An optimum solution is possible even for multi modal objective functions utilizing GAs because they are multi-point search methods. Also, Gas is applicable to discrete search space problems. Thus, GA is not only very easy to use but also a very powerful optimization tool In GA, the search space consists of strings, each of which representing a candidate solution to the problem and are termed as chromosomes. The objective function value of each chromosome is called its fitness value. Population is a set of chromosomes along with their associated fitness. Generations are populations generated in an iteration of the Genetic Algorithm.

The basic genetic algorithm can be explained with the help of following steps:

A. Create

Create a randomly generated initial population (with n number of chromosomes).

B. Evaluate

Evaluate the fitness of each chromosome in the population.

C. Repeat

Repeat the steps given below to create a new population:

1) Selection

Select a pair of chromosomes according to their fitness from the current population.

2) Crossover:

Perform crossover on the pair of chromosomes with a crossover probability to produce new offspring.

3) Mutation:

Mutate the offspring's with mutation probability.

D. Replace

The current population is replaced by new offspring.

E. Check

If the new generation satisfies the end condition then stop, otherwise got to Step C.

III. GENETICOPERATORS

A. Encoding

Encoding of the chromosome is the first step in solving any genetic algorithm. It explains how the genes are represented. It represents the solution in the form of a string.

The different types of encoding based on the type of problem are:

1) Binary Encoding
2) Octal Encoding
3) Hexadecimal Encoding
4) Permutation Encoding
5) Value Encoding
6) Tree Encoding

B. Selection

Selection is the process in which any two individuals (chromosomes) are randomly selected according to their fitness. This means the individual having a high fitness function will have a better chance of being selected. The methods for selection are:

1) Random Selection
2) Tournament Selection
3) Rank Selection
4) Roulette Wheel Selection
5) Proportionate Selection
6) Steady State Selection
7) Stochastic Universal Sampling

C. Crossover

It is also known as recombination. In the Crossover, new solutions are created from the existing population to enrich thepopulation.

Various crossover techniques are:

1) Multi-PointCrossover
2) UniformCrossover
3) ShuffleCrossover
4) Precedence PreservationCrossover
5) Davi's OrderCrossover
6) OrderedCrossover

D. Mutation

Mutation is the process in which one or more gene value of the chromosomes are changed and is send to the next generation. Hence, the new solution may be entirely different from the previous solution. The different operators for mutation are:

1) Interchanging
2) Displacement
3) Insertion
4) Displaced Inversion
5) Random Resetting.

IV. APPLICATIONS

1. SENTIMENT ANALYSIS

Sentiment analysis is the process of analysing the sentiments or views in a given piece of text. Sentiment Classification is a major task in sentiment analysis which deals with classifying the sentiment as a positive sentiment or a negative sentiment. The types of Sentiment Classification are Document Level, Sentence Level and AspectLevel.

The steps involved in sentiment analysis are:

A. Data Collection

Sentiment analysis starts with collecting the data. A proper dataset needs to be defined for analysing and classifying the text in the dataset.

B. Data Pre-processing

After collecting the data, data pre-processing is done this involves removing the unnecessary stop words, repeated words, stemming, removal of emoticons and removal of URLs etc.

C. Feature Selection andExtraction

This is a major step in sentiment analysis. To improve the accuracy, proper selection and extraction of features are very important. Hence, the appropriate feature extraction technique must be chosen for extracting the features.

D. Sentiment Classification

In this phase, various sentiment classification techniques are applied to classify the text. Some popular sentiment classification techniques are Naïve Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (ME),) etc.

E. PolarityDetection

After classifying the sentiments, the polarity of the sentiment is determined. The goal of polarity detection is to decide whether a text expresses positive or negative sentiment.

F. Validation andEvaluation

Finally, results obtained are validated and evaluated to determine the overall accuracy of the techniques used for sentiment analysis.

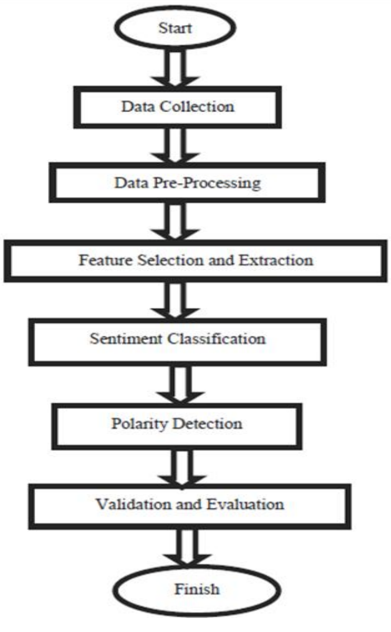

Fig. 2 Sentiment Analysis Framework

A comparative analysis of the various techniques applied to Genetic algorithms in sentiment analysis. Hamidreza Keshvaraz, Mohammad Saniee Abadeh[13] classified tweets into subjective and objective. They performed subjectivity classification on two datasets. Results showed that Genetic algorithm performed better than other baselinemethods.

Dinar Ajeng Kristiyanti, Mochamad Wahyudi [14] used a combined method of feature selection in SVM by comparing Genetic algorithm, Principal Component Analysis (PCA), Particle Swarm Optimization (PSO) for feature selection. The best performance was achieved by PSO 97.00% followed by GA with an accuracy of 94.00%. Lukas Povoda, Radim Burget, Malay Kishore Dutta, Namita Sengar [15] proposed an approach to perform text classification on big data. They used SVM for classification and the Genetic algorithm is used for optimization. The accuracy achieved on Czech texts dataset is90.09%

P. Kalaivani, K.L. Shunmuganathan [16] develop a hybrid genetic algorithm with information gain for feature selection along with K-Nearest Neighbors (KNN) to improve the classification. The accuracy obtained in Movie reviews dataset is 86.00% and on Book review was 80.97%.
Renato S. C. da Rocha, Leonardo Forero, Harold de Mello Jr., Manoela Kohler, Marley Vellasco [18] proposes a two- phase feature selection method. First, an initial preprocessing step is applied, then a feature selection method which is a combination of genetic algorithm and support vector machine is applied on Internet Movie Database (IMBb) reviews. The accuracy achieved is90.0%.

2. GENETIC ALGORITHM BASED FEATURE SELECTION ON PIMA INDIANS DIABETES DATASET

Diabetes Mellitus is a dreadful disease characterized by increased levels of glucose in the blood, termed as the condition of hyperglycaemia.

Goldberg’s Genetic algorithm in the pre-processing stage is used to selects the essential features from the Pima Indians DiabetesDataset.

As a result of feature selection with GA the number of features is reduced to 4 from 8 and the classifier rate is improved to 83.04 %.

As exhaustive search methods are expensive on large feature sets, we choose a stochastic feature selection method. The algorithms initialize the population in the dataset and perform selection, crossover, mutation and termination. The selection process is based on the concept of survival of the fittest.

The formula to calculate the fitness function is given below

$$Fitness\ f(x) = \frac{fitness\ of\ an\ individual\ f(i)}{sum\ of\ fitness\ of\ all\ individuals\ f(I)}$$

The standard pseudo code for the Genetic Feature selection algorithm is asfollows

Algorithm 1.Genetic Feature Selection Algorithm

```
Initialize n = 0
Initialize the individuals in the population as P(n)
Evaluate fitness function for the individuals in the population P(n)
While {termination condition is not satisfied} Do n = n+1 {iterate}
Selection () {pick individuals with better fitness}
Crossover () {combination of parent to form new individuals}
Mutation () {Making changes like bit flip} End while
Return {fittest individuals in the population}
```

Dataset – Pima Indians Diabetes

The Pima Indians diabetes dataset is a publicly available dataset downloaded from UCI machine learning repository.

The dataset comprises of 8 attributes, 768 instances and 1 binary class attribute.

1) Source of the dataset: UCI Machine learning repository
2) About the dataset: This dataset is a collection of health information from Pima Indian women population of 21 years and above in the region of Arizona andPhoenix
3) Attributes:9
a) Number of times pregnant
b) Plasma glucose concentration
c) Diastolic blood pressure mmHg
d) Triceps skin fold thickness(mm)
e) 2-Hour serum insulin (muU/ml)
f) Body mass indexKgm-2
g) Diabetes pedigree function
h) Age in years
i) Binary Class variable

The Classification rate or accuracy of the machine learning algorithm is calculated with the following formula based on the True Positive (TP- Correctly classified as True), False Negative (FN - Wrongly classified as False), True Negative (TN-Wrongly classified as True) and False positive (FP-Correctly classified as False).

The GA Feature selection is applied to the dataset and the classification procedure is repeated on the selected features and the accuracy of prediction is tabulated in Table I.

V. CONCLUSION

Genetic Algorithms proved to be better in finding areas of complex and real world problems. Genetic Algorithms are adaptive to their environments, as this type of method is a platform appearing in the changing environment. In Present these algorithms are more applicable. Several improvements must be made in order that GAs could be more generally applicable. We conclude that in sentiment analysis genetic algorithm is majorly used for optimization to obtain better results. The results obtained from other researchers signify how the genetic algorithm is used in combination with other approaches on different datasets giving different accuracy.

Genetic feature selection algorithm adapted in many state of art literatures is implemented on Pima Indians diabetes dataset and the accuracy is observed.

As a result the proposed method has shown a better performance than existing methods GA provides a comprehensive search methodology for optimization. GA is applicable to both continuous and discrete optimization problems. In global optimization scenarios, GAs often manifests their strengths: efficient, parallelizable search; the ability to evolve solutions with multiple objective criteria; and a characterizable and controllable process of innovation.

REFERENCES

[1] K. G. M. M. Alberti and P. Z. Zimmet, “Definition , Diagnosis and Classification of Diabetes Mellitus and its Complications Part 1 : Diagnosis and Classification of Diabetes Mellitus Provisional Report of a WHO Consultation,” pp. 539–553,1998.

[2] C. Weyer, C. Bogardus, D. M. Mott, and R. E. Pratley, “The natural history of insulin secretory dysfunction and insulin resistance in the pathogenesis of type 2 diabetes mellitus,” vol. 104, no. 6,1999.

[3] J. P. Cunningham and Z. Ghahramani, “Linear Dimensionality Reduction: Survey, Insights, and Generalizations,” vol. 16, pp. 2859– 2900,2014.

[4] D. K. Choubey, S. Paul, S. Kumar, and S. Kumar, “Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute

selection," in Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), 2017, pp.451–455.

[5] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995, pp.338–345.

[6] D. K. Choubey and S. Paul, "GA _ J48graft DT : A Hybrid Intelligent System for Diabetes Disease Diagnosis," vol. 7, no. 5, pp. 135–150,2015.

[7] D. K. Choubey and S. Paul, "GA _ MLP NN : A Hybrid Intelligent System for Diabetes Disease Diagnosis," no. January, pp. 49–59,2016.

[8] F. Jiménez, G. Sánchez, J. M. García, G. Sciavicco, and L. Miralles, "Multi-objective evolutionary feature selection for online sales forecasting," Neurocomputing, vol. 234, no. November 2016, pp. 75–92,2017.

[9] H. R. Kanan and K. Faez, "An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system," Appl. Math. Comput., vol. 205, no. 2, pp. 716–725,2008.

[10] D. Goldberg, "Genetic algorithms in search, optimization, and machine learning, 1989," Read. Addison-Wesley,1989.

[11] F. Jiménez, G. Sánchez, and J. M. Juárez, "Multi-objective evolutionary algorithms for fuzzy classification in survival prediction," Artif. Intell. Med., vol. 60, no. 3, pp. 197–219,2014.

[12] M. A. Hall, "Correlation-based feature selection for machine learning,"1999.

[13] H. Keshavarz and M. S. Abadeh, "SubLex: Generating subjectivity lexicons using genetic algorithm for subjectivity classification of big social data," 1st Conf. Swarm Intell. Evol. Comput. CSIEC 2016 - Proc., pp. 136– 141,2016.

[14] D. A. Kristiyanti and M. Wahyudi, "Feature Selection Based on Genetic Algorithm , Particle Swarm Optimization and Principal Component Analysis for Opinion Mining Cosmetic ProductReview."

[15] L. Povoda and R. Burget, "Genetic Optimization of Big Data Sentiment Analysis," pp. 141–144,2017.

[16] P. Kalaivani and K. L. Shunmuganathan, "An improved K-nearestneighbor algorithm using genetic algorithm for sentiment classification," 2014 Int. Conf. Circuits, Power Comput. Technol. [ICCPCT-2014], pp. 1647–1651,2014.

[17] E. V. Kotelnikov and M. V. Pletneva, "Text sentiment classification based on a genetic algorithm and word and document coclustering," J. Comput. Syst. Sci. Int., vol. 55, no. 1, pp.106–114,2016.

[18] R. S. C. da Rocha, L. Forero, H. de Mello, M. Kohler, and M.Vellasco, "Polarity classification on web-based reviews using Support Vector Machine," 2016 IEEE Lat. Am. Conf. Comput.Intell.,pp.1–6,2.

[19] Noraini Mohd Razali, John Geraghty "A genetic algorithm performance with different selection strategies", Proceedings of the World Congress on Engineering Vol II, 2011.

[20] D. E. Goldberg, "Genetic Algorithm in Search, Optimization and Machine Learning, Reading, MA: Addison-Wesley,1989.