e-PGPathshala

Subject : Computer Science

Paper: Machine Learning

Module: Genetic Algorithms - I

Module No: CS/ML/21

Quadrant I – e-text

Welcome to the e-PG Pathshala Lecture Series on Machine Learning. In this module we will discuss the basics of a very interesting machine learning technique – Genetic algorithms.

Learning Objectives:

The learning objectives of this module are as follows:

- To understand Genetic Algorithms
- To understand the biological nature of Genetic algorithms
- To understand the advantages and disadvantages of Genetic Algorithms

21.1 Introduction

The genetic approach is based on the simulation of the natural law of evolution of a species by Darwin's principle of natural selection process. This approach has been applied to a variety of function optimization problems. It is essentially a search technique used to find true or approximate solutions to optimization and search problems. It is a highly effective technique for searching a large, poorly defined search space even in the presence of complexities such as high-dimensionality, multi-modality, discontinuity and noise. According to Salvatore Mangano in 1995, *"Genetic Algorithms are good at taking large, potentially huge search spaces and navigating them, looking for optimal combinations of things, solutions you might not otherwise find in a lifetime."* Genetic algorithms have been widely used in business, scientific and engineering circles.

21.1.1 History of Genetic Algorithms

It all started with John Holland's work on adaptive systems in 1962. This is the work that laid the foundation for later developments. These developments were further inspired by the book on Adaptation in Natural and Artificial Systems published by Holland along with his students and colleagues in the year 1975. Early to mid-1980s, genetic algorithms began to be applied to a wide range of

problems. In 1992 John Koza used genetic algorithm to evolve programs to perform certain tasks. He chose to call this method "genetic programming" (GP).

21.1.2 Vocabulary

Let us now understand the vocabulary used in the context of Genetic Algorithms (GA).

- Individual - Any possible solution to the problem at hand. An individual is characterized by a set of parameters: Genes
- Population - Group of all individuals or group of all possible solutions
- Trait or Gene- Possible aspect (features) associated with an individual or an individual is characterized by a set of parameters.
- Allele - Possible settings of trait (black, blond, etc.) or possible values a feature can have
- Locus - The position of a gene in the chromosome
- Chromosome – The parameters of the solution (genes) are concatenated to form a string (chromosome)
- Genome - Collection of all chromosomes for an individual.
- Genotype - encoded representation of individual. The genotype contains all information to construct an organism: the phenotype
- Phenotype - decoded representation of individual
- Reproduction - process on the chromosome of the genotype
- Fitness - measured in the real world (‘struggle for life’) of the phenotype
- Mapping – is the process of decoding of the phenotype
- Mutation - variability operator that modifies a genotype
- Recombination/Crossover - variability operator that mixes the genotypes
- Fitness - performance of a phenotype with regard to an objective
- Iteration – the number of generations used

21.2 Use of GA

Genetic algorithms are best suited for applications where alternate solutions are too slow or too complicated. It is used in situations where there is a need of an exploratory tool to examine new approaches. Genetic algorithms are generally

used when we have a problem that have characteristics similar to problems that have already been successfully tackled using genetic algorithms. Genetic algorithms have been used along with other approaches in a hybrid manner to solve problems since it is to combine it with other methods. Genetic algorithms are used when the advantages provided by the genetic algorithms meet the key requirements of the problem that has to be solved.

21.3 Applications of GA

Genetic algorithms have been used in many domains and for many applications in those domains. GAs have been used in many control type of applications. These include problems such as control of gas pipeline, pole balancing, missile evasion and pursuit of objects. GAs have also been used for optimization of parameters in a design scenario such as in the design of semiconductor layout, aircraft design, keyboard configuration and filter design of signal processing, etc. These algorithms have been used for trajectory planning in robotics and in game playing for games such as poker, checkers, prisoner's dilemma, etc. GA's are suitable for combinatorial optimization problems such as set covering, travelling salesman, routing, bin packing and graph coloring and partitioning.

21.4 Biological Background

As we have already discussed genetic algorithms have been inspired by natural evolution. We have a population of individuals where a particular individual is a feasible solution to problem. Each individual is characterized by a fitness function where the individual with a higher fitness is a better solution. Basically depending on their fitness, parents are selected to reproduce offspring for a new generation where the fitter individuals have more chance to reproduce and the offspring so formed has a combination of properties of the two parents. The new generation formed has the same size as the old generation. If well designed, the population will converge to optimal solution. Now let us understand the biological background of each of the components of the genetic approach.

21.4.1 Biological Background of the Cell

Every animal cell is a complex system consisting of many small "factories" working together. The nucleus is at the center of the cell. It is this nucleus that contains the genetic information.

21.4.2 Biological Background of the Chromosomes

Genetic information is stored in the chromosomes where each chromosome consists of DNA strands. In humans, chromosomes form pairs and there are 23 such pairs. The chromosome is divided into parts called genes which actually code the properties. Every gene has an unique position on the chromosome

which is it's locus. The posibilities of the genes for one property or trait is called the allele.

•

21.4.3 Biological Background of Genetics

The entire combination of genes is called the genotype. The genotype will eventually develop to form a phenotype. The alleles or the traits can be either dominant or recessive. Dominant alleles will always express from the genotype to the phenotype.However recessive alleles can survive in the population for many generations, without being expressed.

21.4.4 Biological Background of "Reproduction"

The reproduction of genetical information is through two processes namely mitosis and meiosis. Mitosis is the process of copying the same genetic information to new offspring in other words there is no exchange of information. It is the normal way of growing of multicell structures, like organs. On the other hand, meiosis is the basis of sexual reproduction. After meiotic division 2 gametes appear in the process. In reproduction two gametes conjugate to a zygote which will now become the new individual. Hence genetic information is shared between the parents in order to create new offspring. During reproduction "errors" can occur and it is due to these "errors" that genetic variation exists. The most important "errors" are recombination (cross-over) and mutation which we will discuss in detail later.

21.4.5 Biological Background of "Natural selection"

The origin of any species is essentially "Preservation of favourable variations and rejection of unfavourable variations." There are more individuals born than can survive, so there is a continuous struggle for life. Individuals with an advantage have a greater chance to survive that is so the so called survival of the fittest. Some of the important aspects in natural selection are the issues of adaptation to the environment and the isolation of populations in different groups which cannot mutually mate. If small changes in the genotypes of individuals are expressed easily, especially in small populations, we speak of the concept of genetic drift

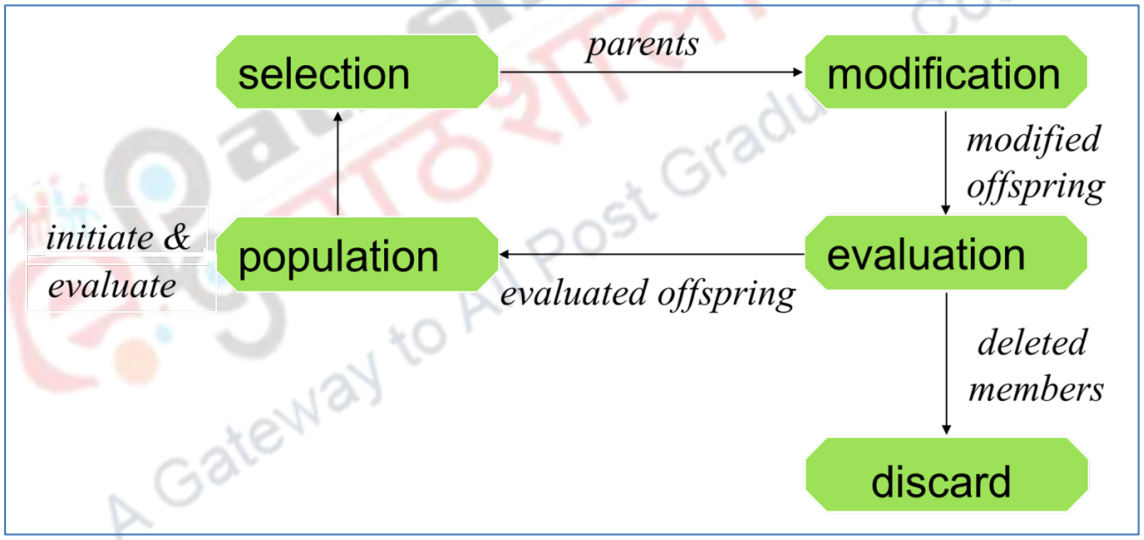
21.5 The Metaphor

Now let us understand the metaphor between the biological background and the components of the genetic algorithm. When we have an optimization problem to solve this is equivalent to the environment. The feasible solutions are the possible solutions in that environment. The first step is the encoding

technique. This is the process of representing the solution in the form of a string that conveys the necessary information. Similar to the chromosome, each gene controls a particular characteristic of the individual, similarly, each bit in the string represents a characteristic of the solution. All kind of alphabets can be used for a chromosome (numbers, characters), but generally a binary alphabet is used.

The initialization procedure is the creation process. The evaluation function is based on the environment. The reproduction is based on the selection of parents where the genetic operators like mutation and recombination are important. However the parameter settings are a question of practice and art. Finally the quality of the solutions indicates the individual's degree of adaption to its surrounding environment. The computer model introduces simplifications relative to the real biological mechanisms, however surprisingly complex and interesting structures have emerged that help solve some complex optimization problems.

21.6 Basic Genetic Algorithm



Initially many individual solutions are randomly generated to form an initial population. The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Traditionally, the population is generated randomly, covering the entire range of possible solutions in the search space. Sometimes in order to allow search in spaces where optimal solutions are likely to be found, the solutions are seeded.

Initially we start with a subset of *n* randomly generated possible solutions to a problem from the search space (i.e. chromosomes). This is the population used to produce a next generation of individuals by reproduction. Individuals with a higher fitness have more chance to reproduce (i.e. natural selection). The primary objective of the recombination operator is to emphasize the good

solutions and eliminate the bad solutions in a population, while keeping the population size constant. We need to repeatedly do the following, evaluate each of the attempted solutions, keep a subset of these solutions (the “best” ones) and then use these solutions to generate a new population and finally quit when you have a satisfactory solution or you have run out of time.

21.7 Design Decisions

GAs have high flexibility and adaptability because of the way in which the problem is represented, the genetic operators and the associated parameters used, the mechanism of selection, the size of the population selected and the actual fitness function used. Please note that decisions are highly problem dependent, and that the parameters not independent in that they cannot be optimized one by one The parameters have to be searched so as to find a balance between exploration that is look for solutions in new search regions and exploitation that is carry out exhaustive search in the current region itself. Parameters can be made adaptable from values being high in the beginning that is going for exploration to values being low that is going for exploitation, or the parameters may even be subject to evolution themselves. The balance between exploration and exploitation is influenced by both mutation and recombination. Diversity can be achieved by creating individuals that are in new regions. We can also fine tune the solutions in current regions by the process of selection that focus on interesting regions.

When we design the genetic approach we need to consider some factors. We first need to create an initial population that has a lot of diversity. In general, during search we need to choose areas that have previously proven to be good; this in essence means that we are sacrificing diversity over the iterations. However this approach at premature convergence that the quick loss of diversity poses high risk of getting stuck in a local optima.

For the evolvable nature of the algorithm to be effective the fitness landscape should not be too rugged, the traits must be of heredity nature and small genetic changes should be mapped to small phenotype changes.

21.8 GA Operators

21.8.1 Selection

During each successive generation, a proportion of the existing population is selected to breed a new generation. The main idea during selection is to focus on fittest individuals. Individual solutions are selected based on fitness function values, where fitter solutions are typically more likely to be selected. Generally selection methods preferentially select the best solutions. However sometimes other methods rate only a random sample of the population, since rating all the solutions in the search space may be a very time-consuming process. Most of

the fitness functions are stochastic that ensure that only a small proportion of less fit solutions are selected. This helps keep the diversity of the population large, preventing premature convergence on poor solutions. Selection methods including roulette wheel selection and tournament selection will be discussed in the next module.

21.8.2 Reproduction – Crossover and Mutation Operators

Genetic algorithms through the process of reproduction produce new generations of improved solutions by selecting parents with higher fitness ratings or by at least a greater probability for such parents to being contributors and by using random selection.

The next step of genetic algorithms is to generate a second generation population of solutions from those selected through genetic operators such as crossover (also called recombination), and/or mutation. Recombination through the process of cross over decomposes two distinct solutions and then randomly mixes their parts to form novel solutions and thereby add alternative solutions to population. In other words, crossover means choosing a random position in the string (chromosome) and exchanging the segments either to the right or to the left of this point with another string partitioned similarly to produce two new off spring.

After selection and crossover, you now have a new population full of individuals. Some are directly copied, and others are produced by crossover. In order to ensure that the individuals are not all exactly the same, you allow for a small chance of mutation. This operation is performed to make sure that most of the search space is covered. The mutation operator induces a change in the solution, so as to maintain diversity in the population and prevent premature convergence.

For each new solution to be produced, a pair of "parent" solutions is selected for breeding from the pool selected previously. By producing a "child" solution using the above methods of crossover and mutation, a new solution is created which typically shares many of the characteristics of its "parents". New parents are selected for each child, and the process continues until a new population of solutions of appropriate size is generated. These processes ultimately result in the next generation population of chromosomes that is different from the initial generation.

Generally the average fitness will have increased by this procedure for the population, since only the best organisms from the first generation are selected for breeding, along with a small proportion of less fit solutions, for reasons already mentioned above.

21.9 Convergence and Diversity

Convergence and premature convergence are issues that are associated with genetic algorithms. GA continues while fitter chromosome exist in population. The population will converge so that all individuals are as fit as each other. The converged population will not evolve further if the best solution found. However though this process actually converges prematurely, the algorithm terminates.

Diversity needs to be maintained in genes. Individuals with the same fitness might keep different gene strings. There are chances that some genes may survive if environment changes. The chance to break premature convergence is by using mutation operators.

21.10 Phenotype and Genotype

Phenotype describes what an individual looks like, after fitness function calculation. Genotype is used to describe genes on chromosome. Individuals with the same phenotype may have totally different genotypes. Genotype maintain diversity for chance of changes. Population with the same genotype individuals can still be evolved using the mutation operator. If the mutation operator is too strong then too much diversity will be introduced in genotype and then no population would be able to converge. If the mutation is too weak there is less chance to change and premature convergence becomes easy.

21.11 Exploitation and Exploration

Exploitation in general takes the current search information from the experience of the last search to guide the search toward the direction that might be close to the best solutions. That is in other words exploitation allows focusing in areas of more promising solutions. Exploitation is achieved through selection operator and crossover operator.

Exploration on the other hand widens the search to reach all possible solutions around the search space. In other words exploration allows the entire search space to be explored. Exploration is achieved through mutation operator and crossover operator.

An important aspect of the genetic approach is the balancing between exploitation and exploration. High exploitation leads to premature convergence and high exploration results in non-convergence and sometimes may not lead to a fitter solution.

21.12 Benefits of GAs

Basically the concept of genetic approach is easy to understand. It is modular and separate from application. It supports the important and complex problem of multi-objective optimization. GA is for “noisy” environments. Genetic

approach always finds an answer and the answer gets better with time. The approach is inherently parallel and easily distributed. There are many ways to speed up and improve a GA-based application as more knowledge about problem domain is gained. Genetic approach makes it easy to exploit previous or alternate solutions and forms flexible building blocks for hybrid applications. It has history and has been applied to many optimization problems.

21.13 Disadvantage of GAs

One of the main disadvantages of genetic algorithms is that no guarantee for optimal solution within a finite time or in other words they lack the killer instinct. It has comparatively weak theoretical basis. In order to arrive at a solution there may be a need for extensive parameter tuning since the search capability is sensitive to parameter settings and operators adopted. The representation, fitness function, and even the position of bits influence judgment of the achievement of fitness optimum that is whether the optimum is global or only local. Genetic algorithms are often computationally expensive and slow to reach optimum solutions.

21.14 Advances in Genetic Approaches:

There are more recent inputs from biology that contribute to new directions in genetic approach. Populations can be spatial, e.g. for “speciation” where interaction that is mating and competition is localized to maintain diversity. Populations can have structure, e.g. niche protection where the competition will Diploidy and dominance have been used in genetic algorithms to improve performance in time-varying optimization problems. Diploidy is a concept to increase diversity where recessive genes are allowed to survive in a population though they are inactive. When changes in the environment are more conducive these recessive genes become active. In the cross over operation N-point crossover and other variants are also being suggested. The concept of morphogenesis instead of simple function mapping that is allowing for modularity and making crossover less fatal.

Summary

- Discussed Genetic Algorithms
- Discussed the biological nature of Genetic algorithms
- Discussed the advantages and disadvantages of Genetic Algorithms