Slide 0

Genetic Programming II

Reference book: "Genetic Programming An Introduction" by W. Banzhaf (Morgan Kaufmann Publishers)

GP features

Slide 1

- automatic programming must
  - produce an entity that runs on a computer
  - solve a broad variety of problems
  - require a minimum of user-supplied problem-specific information
  - not require the user to prespecify the size and shape of the ultimate solution
  - implement all the familiar and useful programming constructs
- GP is fundamentally different from other approaches in terms of
  - its representation (namely, programs)
  - the role of knowledge (none)
  - the role of logic (none)
  - its mechanism for getting to a solution within the space of possible solutions

GP features (Cont'd)

Slide 2

- practical application areas where
  - the interrelationships among the relevant variables are poorly understood
  - finding the size and shape of the ultimate solution is a major part of the problem
  - conventional mathematical analysis does not or cannot provide analytic solutions
  - an approximate solution is acceptable
  - small improvements in performance are routinely measured and highly prized
  - there is a large amount of data that requires examination, classification, and integration such as
    - molecular biology for protein and DNA sequences
    - astronomical data
    - satellite observation data
    - financial data
    - marketing transaction data
    - World Wide Web data

machine learning

Slide 3

- a process that
  - begins with identification of the learning domain
  - ends with testing and using the results of learning
- key parts
  - the learning system
  - learning domain
  - training set
  - testing the results of the learning process

machine learning (Cont'd)

Slide 4

- learning domain
  - any problem
  - set of facts to identify "features" of the domain
  - a result or results
  - for example
    - learning domain: stock market
    - features (input): closing S&P index for the past 30 days
    - classes (output): the closing S&P index tomorrow
- training sets, training data
  - specific past examples from the learning domain
  - instances of the relationship between *the chosen features and the classes*
  - called training cases, training instances →all training instances (training set)
  - fitness cases in GP

machine learning (Cont'd)

Slide 5

- training
  - machine learning occurs by training
  - attempts to learn from the training set
  - in GP
    - GP must learn a computer program
    - that can predict the outputs of the training set from the inputs
- major issues in machine learning
  - learning algorithms
    - GP use a learning algorithm based on an analogy with natural evolution
    - neural networks are based on an analogy with biological nervous systems
    - Bayes/Parzen systems are based on statistics

machine learning (Cont'd)

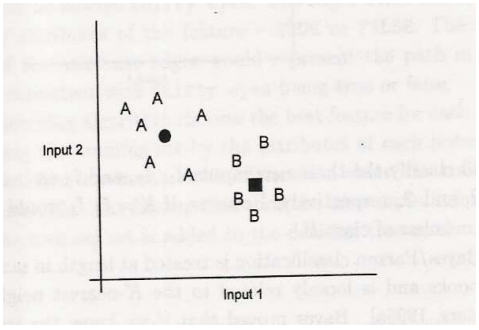
Slide 6

- classification of learning algorithms
  - how are solutions represented
  - what search operators does the learning algorithm use to move in the solution space
  - what type of search is conducted
  - is the learning supervised or unsupervised
- representing the problem
  - problem representation
    - define the set of all possible solutions
    - called candidate solutions
    - define the space of candidate solutions
  - three different levels
    - representation of the input and output set
    - representation of the set of concepts the computer may learn
    - interpretation of the learned concepts as outputs

machine learning (Cont'd)

Slide 7

- three types of representation
  - Boolean representation
  - threshold representation: more powerful than Boolean
  - case-based representation
    - instance averaging

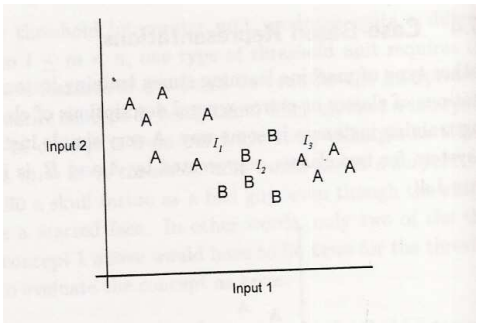


→difficult to linearly non-separable classes

machine learning (Cont'd)

Slide 8

- K-nearest neighbor approach
  →when K=3, look at the 3 nearest neighbors, if two of them are from class A, then class A



- tree representation
- genetic representation

machine learning (Cont'd)

Slide 9

* GP is capable of evolving any solution
  - may include Boolean operators, Boolean representations
  - can implement a threshold function using IF/THEN structure
  - can implement decision trees using IF/THEN or SWITCH
  - can implement a case-based system

☛ GP is a superset of other ML representation

machine learning (Cont'd)

Slide 10

- transforming solutions with search operators
  - what is it?
    - a problem →candidate solutions (by representation)
      →huge candidate solutions for non-trivial problems
      →impossible or impractical to evaluate the entire space
    - must define how it will search a limited portion
    - which candidate solution will it evaluate first, which next, and next, and when will it stop?
  - search operators
    - define how an ML system chooses solutions to test and in what order
    - search operators or transformation operators define the area of the representation space that actually will be searched

machine learning (Cont'd)

Slide 11

- three types of search operators
  - generality/specificity operators
    a search from the most general to the most specific solutions
  - gradient descent operators
    in neural networks, the weights are adjusted according to a gradient descent algorithm
  - genetic programming operators
    primary transformation operators: crossover and mutation
    predominant operator: crossover (the basis of the GP building block hypothesis)
    building block hyperthesis: useful building blocks are accumulated for better solution by crossover
- GP
  - can implement generality/specificity operators by automatically defined functions or genetic operators

machine learning (Cont'd)

Slide 12

- the strategy of the search
  - three types of search
    - blind search
    - hill climbing
    - beam search
  - blind search
    - picking a solution using no
      - information about the problem or
      - results from previous steps in the search
    - in tree representation, breadth-first and depth-first tree search exist (exhaustive search)
    - only for small search space

machine learning (Cont'd)

Slide 13

  - hill climbing
    - operation
      - start in one spot in the search space
      - transform that solution
      - keep the new solution if it is a better
      - otherwise, it is discarded and transforms the original solution again
      - until a termination criterion is met
    - simulated annealing (SA) and many neural network training algorithms are typical
    - only one solution is considered at a time and only one path through the solution space is explored
  - beam search
    - GA, GP, and beam search maintain a population of search points
    - a compromise between exhaustive search and hill climbing
    - GP is a form of beam search in that
      - ML evaluation metric for the beam is called the fitness function in GP
      - the beam of machine learning is referred to as the population in GP

machine learning (Cont'd)

Slide 14

- learning
  - three types of learning
    - supervised learning
      - when correct training instances are given
      - many GP application use supervised learning
      - the fitness function compares the output of the program with the desired result
    - unsupervised learning
      - when no correct outputs are given, the system looks for patterns in the input data
      - Kohonen SOFM is a good example
      - GP is not normally used for unsupervised training, but it would be possible

machine learning (Cont'd)

Slide 15

    - reinforcement learning
      - fall between supervised and unsupervised learning
      - although correct outputs are not specified, a general signal for quality of an output is fed back to the learning algorithm
      - many of the fitness functions in GP are more complex than just comparing the program output to the desired output

      this could be considered as reinforcement learning systems

- summary
  - GP represents a problem as the set of all possible computer programs
  - GP uses crossover and mutation as the transformation operators to change candidate solutions into new candidate solutions
  - GP uses a beam search
    - where the population size contributes the size of the beam
    - where the fitness function serves as the evaluation metric
  - GP typically is implemented as a form of supervised machine learning, but it is perfectly possible to use GP as a reinforcement or an unsupervised learning system

Genetic programming and biology

Slide 16

- DNA base pairs
  - regularly store informations learned through biological evolution
  - sequences of DNA act like instructions or partial instructions in computer programs
- minimal requirements for evolution to occur
  - four essential preconditions
    - reproduction of individuals in the population
    - variation that affects the likelihood of survival of individuals
    - heredity in reproduction
    - finite resources causing competition

Genetic programming and biology (Cont'd)

Slide 17

- the genetic code-DNA as a computer program
  - DNA may be regarded as a complex set of instructions for creating an organism
  - four bases of DNA
    - (A)denine, (G)uanine, (C)ytosine, (T)hymine
    - base parings: A $\leftrightarrow$T, T $\leftrightarrow$A, G $\leftrightarrow$C, C $\leftrightarrow$G
  - four bases pairs bond to each other forming a ladder
    - form a double helix
    - provide 3-dimensional properties
    - give a redundancy that repair mechanisms based on (only one of the two strands reconstructs the entire DNA)

Genetic programming and biology (Cont'd)

Slide 18

- codons and amino acid synthesis
  - codons
    - three base pairs
    - a template for the production of a particular amino acid or a sequence termination codon
    - for examples
      - ATG codes for methionine
      - CAA codes for glutamine
      - CAG also codes for glutamine
    - $4^3 = 64$ codons possible, but only 20 amino acids exist
    - different codons code the same amino acid (☛redundancy)
    - redundancy
      - if one codon is mutated to another codon that produces the same amino acid
      - then the protein is not changed
  - DNA instruction sequencing
    - provide informations to synthesize proteins in organism
    - from 5 prime to 3 prime site

Genetic programming and biology (Cont'd)

Slide 19

- polypeptide, protein, and RNA synthesis
  - DNA acts to manufacture polypeptides, proteins, and non-translated RNA (tRNA and rRNA)
  - proteins are complex organic molecules that are made up of many amino acids
  - polypeptides are protein fragments
  - DNA transcribes RNA molecules, which then translate into one or more proteins or polypeptides
- genes and alleles
  - oversimplified explanation
    - a gene is a location on the DNA that decides what color your eyes will be
    - slight variations make your eyes green or brown
    - the variations are called "alleles"

Genetic programming and biology (Cont'd)

Slide 20

- biologists explanation
  - ∗ trait
    - · adjacent sequences of DNA do act together to affect specific traits
    - · but, a single gene can affect more than one trait
    - · moreover, widely scattered DNA may affect the same trait
  - ∗ junk DNA
    - · the transcriptional portion of DNA are separated by long meaningless DNA (junk DNA)
    - · junk DNA does not qualify as a gene
  - ∗ exon and intron
    - · intron (junk DNA) are removed from the RNA →mRNA
    - · mRNA translation →protein
    - · transcriptional portion is called exon
    - · intron
      - play a role in preventing damage to exons during recombination
      - provide shuffling and combining slightly different variations of the functional parts

Genetic programming and biology (Cont'd)

Slide 21

- genomes, phenomes, and ontogeny
  - genotype and phenotype
    - ∗ genotype
      - · the genome or genotype of an organism is the DNA
      - · the genome is the principal mechanism for variance by mutation and recombination
    - ∗ phenotype
      - · the phenome or phenotype is the set of observable properties and the behavior of the organism
      - · natural selection acts on the phenotype (not on the genotype)
  - ontogeny
    - ∗ the development of the organism from fertilization to maturity
    - ∗ the link between the genotype, the phenotype, and the environment
    - ∗ one-way street: changes of DNA can change the organism, not vice versa
    - ☛ heredity takes place at the genotype, natural selection acts only at the phenotype

Genetic programming and biology (Cont'd)

Slide 22

- stability and variability of genetic transmission
  - for evolution to occur, genetic transmission must be stable (for good traits) and variable (for new traits)
  - stability
    - redundancy
    - repair
    - homologous sexual recombination
      - prevent the fixing of negative mutations
      - provide a chance to repair of damaged DNA
  - genetic variability
    - mutation
      - changes from one base pair to another
      - additions or deletions of one or more base pairs
      - DNA sequence rearrangements

Genetic programming and biology (Cont'd)

Slide 23

  - homologous recombination
    - two requirements
      - two identical or almost identical DNA segments
      - if the two DNA segments to be exchanged can be matched up so that the swap point is functionally identical

Genetic programming and biology (Cont'd)

Slide 24