Genetic Algorithms

GA in Clustering

The task of the GA is to find the appropriate cluster centers such that the clustering metric for finding minimum distances is minimized.

Representation

Each string is a sequence of real numbers representing the K cluster centers. The length is N*K for an N dimensional space, where the first N positions are for the first cluster, and so on.

Fitness Computation

1. Clusters are formed according to the centres encoded
2. Ties are resolved arbitrarily
3. New centroids are computed again and the chromosome is updated
4. Fitness score is calculated as 1/M where M is the clustering metric.

Selection

Choose 50% of the top fit chromosomes and the other 50% by crossover.

Crossover

A single point crossover with a fixed crossover probability can be used. Generate the crossover point for a length l. The portions lying to the right of the crossover point are exchanged.

Mutation

For binary genes, we can just flip a random number, or select a number d in the range [0,1] and update the gene at position v as v = v +/- 2dv if v!=0 and v +/- 2d otherwise.

Termination

We can terminate after a fixed number of iterations. Maintain the highest scores somewhere for each hypothesis and output the best ones. Inter-cluster distance can also be used for the fitness computation.

GA in Decision Trees - Approach 1

Representation

Trees are generated randomly, where growth of a tree is dictated by a parameter that indicates the maximum tree depth, but no less than 2 levels.

The GA starts with a root node and 2 children, and with some probability p, it decides whether the children are split or the node becomes a leaf. For a leaf, a random class label is assigned.

The trees are hence, represented as arrays.

Fitness Calculation

We consider depth and accuracy to calculate the fitness with some relative weights as

$$F = \alpha_1 f_1 + \alpha_2 f_2$$

where

$$f_1 = 1 - \frac{Correctly\ classif ied}{Total\ samples}$$

$$f_2 = \frac{Current\ depth}{Target\ depth}$$

Selection

50% selection with any algorithm of choosing

Crossover

Select a random node, and identify the subtrees in the parents. A new individual is formed by replacing the subtree from the first parent by the one in the second parent. Some validation is done to ensure the formed tree is correct.

Mutation

Change a node and branch values at random.

Termination

Terminate if

1. Fixed iterations
2. Fitness does not increase

GA to learn parameters of an ANN

Representation

Each chromosome is a string of all the weights as a vector.

Each solution in the population has a 1D vector for the GA and a matrix for the ANN.

Fitness Function

Fitness is the accuracy of the ANN. This eliminates the need to perform backpropagation, eliminating vanishing or exploding gradients. Only forward propagation is done to get the output.

Disadvantages

If number of parameters to be learnt are too many, large initial population will be required with too many generations.