A(E,D)4M33BIA: Individual project

Task is to

1. Design an evolutionary algorithm for chosen problem.
2. Implement and experimentally evaluate the proposed algorithm.
3. Write a report consisting of two parts
   a) Specification of chosen solution representation, selection strategy, crossover and mutation operators, evolutionary model, and definition of the fitness function.
   b) Report on the performance observed with the implementation.

      Use
      - tables to present statistics such as the absolute best solution values, the mean best-of-run values (accompanied by standard deviation), mean computation time etc. and
      - graphs showing mean/typical convergence curves.

      Discuss the achieved results.

List of topics:

1. Genome Sequence Analysis

Given: A set of fragmentary gene segments. Only four letters will be used: A (adenine), C (cytosine), G (guanine), and T (thymine). The segments have the same length. For example, an input could look like:

TCGG
GCAG
ATCG
CAGC
GATC

Goal is to reconstruct the gene sequence i.e. to reassemble the fragments into the shortest possible gene that can be made from these pieces. The rules for assembling the final sequence are:

1. The sequence must use all the segments (and only the segments) provided
2. You cannot flip the segments.
3. The best answer is the shortest answer.

Example: CAGCAGATCGG and GATCGGCAGC are possible final sequences of length 11 and 10, respectively.

Hint: Since this is a permutation problem, you might want to try some of the crossover operators designed for the TSP, such as the Partially Mapped Crossover or Edge-Recombination Crossover.

2. Design of Molecule

You are trying to model the structure of a molecule (which for the sake of simplicity will be limited to two dimensions).

Given: A collection of plastic connecting rods that represent the distance between pairs of atoms in the molecule, given in a symmetric distance matrix. A -1 in the [i,j] location signifies that there is no connection specified between the two atoms in question.

Goal is to fit the connecting rods together into a consistent two dimensional shape so that the actual lengths between atoms follow as precisely as possible the preferred ones. It is not always possible to get 100% precise solution, so some of the lengths you specify will be slightly longer or shorter than their preferred length.

This isn't always possible, so some of the lengths you specify will be slightly longer or shorter than their preferred length. **Stretching** or **compressing** the connecting rods is legal, but it causes a strain. The strain on the rod between any two atoms is defined as the absolute value of the difference between the preferred distance and the actual distance between atoms *i* and *j*. The sum of strain values over all rods should be minimized.

Focus:

1. Binary vs. real representation and appropriate genetic operators.

3. Gerrymandering

Gerrymandering is the process of carving up an electoral district into strange shapes in order to derive a political advantage. You are given the task of preparing for an upcoming election in the state of Rectanglia. As the director of redistricting, your job is to divide the state into N districts of equal population.

Given: A matrix A in which each element corresponds to the population in a given square mile.

Goal is to find a matrix B that indicates which voting district each square mile belongs in and voting districts must be contiguous, or connected in the four-neighbor sense (no diagonal connections).

Example. Suppose, given the following census data for Rectanglia in matrix A, you are told to divide the state into N = 3 districts of equal population.

$$A = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 12 & 11 & 2 & 12 \\ 40 & 28 & 10 & 2 \end{bmatrix}$$

Here's one way to optimally divide the state.

$$B = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 3 & 2 & 2 & 1 \end{bmatrix}$$

Focus: Design of crossover and mutation operators that produce only valid solutions.

4. Optimal allocation of new water reservoirs

Given: A map of *M* cities, each defined by its coordinates [*x,y*].

Goal is to build new reservoirs in *N*<*M* cities so that the total length of the pipelines connecting all cities with its nearest reservoir is minimized.

Focus:

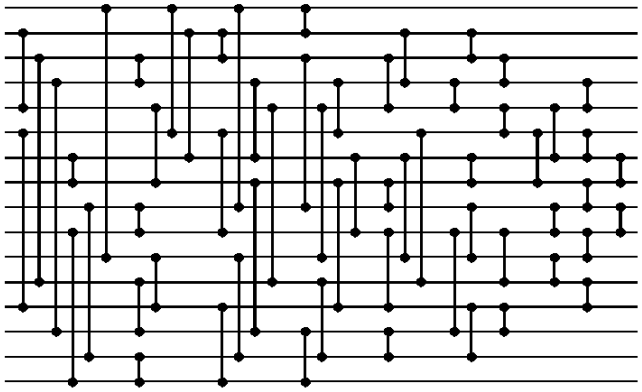
1. Try to use different metrics (euclid, manhattan, ...) for calculating the pipe length.
2. Reservoirs must be positioned in one of the cities or reservoirs can be at arbitrary positions within the map.

5. Sorting networks (Wikipedia)

Given: A sequence of *N* numbers and a set of comparators, where each comparator connects two wires and sorts the values by outputting the smaller value to one wire, and a larger value to the other.

Goal is to design for the given sequence of numbers a sorting network with minimal number of comparators.

Example:



Each vertical line represents a comparator that switches the input numbers if they are not in desired order.

6. Piecewise linear approximation of a set of points

Given: A set of points [*x*,*y*].

Goal is to find a piecewise linear curve that minimizes the chosen objective (e.g. minimal sum of squared deviations or sum of absolute deviations).
The final approximation curve must be continuous.

Focus:

1. Input parameter is the desired number of linear segments.
   The number of lines is not defined – then the number of linear segments needed to sufficiently approximate the data will be minimized.
2. Representation
   - Set of end-points of partial linear segments.
   - Set of lines.

7. Magic squares

Given: A matrix *N* x *N* and a set of numbers from interval <1, N^2>.

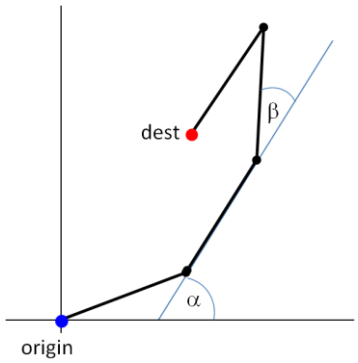
Goal is to place the number into the table so that every number is used just once and the numbers in all rows, all columns, and both diagonals sum to the same constant.

8. Robotic arm with multiple joints

Given: A robotic arm consisting of a number of segments connected by 2-degree of freedom joints. All segments are of the same length. The first joint is connected to the origin at <0,0>. The target position is at <x,y>.

Goal is to set angles of the joints so that the robotic arm reaches the target position.

Parameters of the problem are the lengths of the segments and the numbers of segments.



Focus:

1. Representation – use absolute (α) vs. relative (β) angles.

9. Extending tram line system

Given: A map of the city with existing tram stops and the system of tram lines. Vertices represent the tram stops and the edges represent direct connection between two stops. Every edge is assigned a value specifying the the time needed to walk the way from one station to the other one. If the tram is used then the time reduces *k*-times.

Goal is to design a new line of length *m* so that the sum of travel times from a given node *v* to all other stops is minimized. The new line must be connected and no loops nor branches are allowed.

Focus: Design of crossover and mutation operators that produce only valid solutions.

10. Maximum Satisfiability problem (MAX-SAT)

Given: A set of logical formulas composed of boolean variables x_1, x_2, ..., x_N. Formulas are in *conjunctive normal form* (CNF) $C_1 \wedge C_2 \wedge \ldots \wedge C_m$ such that

1. a clause is a disjunction (OR) of literals, such as (x5 or not(x12)),
2. NOT operators are only applied directly to variables, such as not(x12),
3. a formula is a conjunction (AND) of clauses, such as (x5 or not(x12)) and (x3 or not(x7) or x11).

Goal is to assign logical values to all variables in such a way that the number of clauses that are satisfied (i.e. they evaluate to TRUE) is maximized.

Focus:

1. Different crossover operators – 1-point, 2-point and uniform crossover.
2. Different replacement strategies – generational, steady-state.

11. Allocation of modules to processors

Given: A set of N modules; each module has assigned its computational complexity. A table of intermodule communication C, where position $C[i,j]$ represents a communication rate between modules i and j.

Goal is to find an allocation of modules to M processors such that the overall computational load is evenly distributed to processors and the interprocessor communication is minimized.

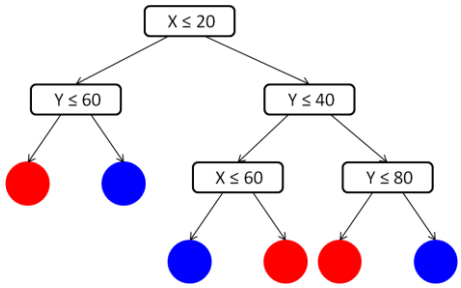
Focus:

1. Multi-objective approach – NSGA-II, SPEA2.
2. Single-objective with weighted individual objectives.

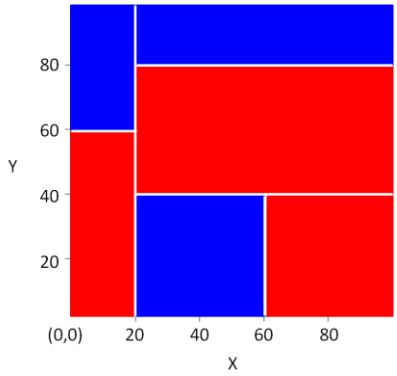
12. Classification problem

Given: A set of training points $T=\{(x_i,y_i)\colon i = 1..n\}$ that is split into two disjunctive sets of positive ***P*** and negative ***N*** cases.

Goal is to find a decision tree classifying as much points as possible correctly. At each level the classified space is divided into 2 subspaces in one of the dimensions only. Sublevels continue dividing.



a) Decision tree



b) splitted decision space

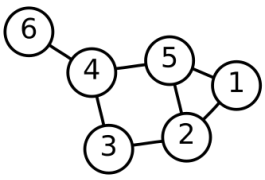
Hint: Use genetic programming to evolve the DT.

Focus:

1. Fitness takes into account only the classification accuracy.
2. Fitness takes into account both the classification accuracy plus the complexity of the DT.

13. Vertex cover problem

Given: An undirected graph with N vertices and M edges given by a symmetric binary matrix $G[N\times N]$, where $G(i, j) = 1$, iff there is an edge between i and j.



Vertex cover is a subset V' of the vertices of the graph which contains at least one of the two endpoints of each edge.

Goal is to find a vetrex cover of minimum size.

Example: vc1={1,3,5,6}, vc2={2,4,5}

14. Approximating euclidean distance

Motivation: The euclidean distance must be computed in many tasks. However, it is not cheap operation since it requires a square root. There are many efforts to use approximations of euclidean distance which are cheaper, but not precise.

One of the approaches is to use the following approximation for D-dimensional space:

$$d_1(x_1, x_2) = a_1 \left|\Delta_{\pi(1)}\right| + a_2 \left|\Delta_{\pi(2)}\right| + \ldots + a_D \left|\Delta_{\pi(D)}\right|$$

where

$$\Delta_d = x_{1,d} - x_{2,d}$$

is the distance of the points in the d-th coordinate, and

$$\pi$$

is a permutation which sorts the deltas, so that

$$\left|\Delta_{\pi(1)}\right| \geq \left|\Delta_{\pi(2)}\right| \geq \ldots \geq \left|\Delta_{\pi(D)}\right|.$$

Goal: Using any instance of evolutionary algorithm, the goal is to find values of individual coefficients a_i that would result in an approximation that is as close as possible to the right values of distance.

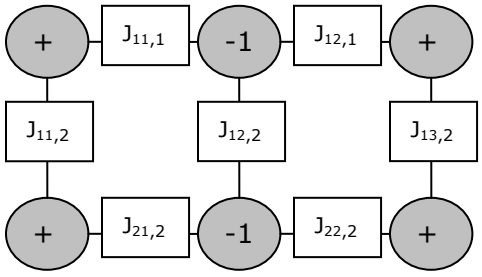
Hint:

- Choose a set of points in a grid that will serve for the evaluation.
- Try to find coefficients *a* at least for 2- and 3- dimensional spaces.
- Try to use various reasonable measures of quality of the approximation (Average error? Maximal error? Another measure?).
- Display graphically the differences between real Euclidean distance and your approximation.

15. Ground state of spin glass model

Given: A set of spins with possible positions -1 and +1. The spins are situated in a 2-D rectangular grid. The nearest neighbors *i*, *j* are connected with linkages whose strengths are defined as coupling constants J_{ij}. Energy of the model is

$$H = -\sum_{i,j} J_{i,j} s_i s_j$$



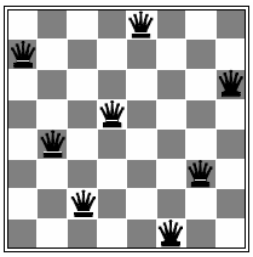
The goal is to implement and test a suitable evolutionary method that finds the state (set of spin orientations) with minimum energy (ground state). Consider just 2D model with $M \times N$ spins. The inputs will be the values of $J_{i,j}$. To prove the

functionality, the method must be able to find exact solution of Ising ferromagnet, a special case of spin glass model, where $J_{i,j} = J>0$.

Focus:

1. Different crossover operators – 1-point, 2-point and uniform crossover.

16. N-queen Problem



Given: A chessboard *N* x *N* and *N* queens.

Goal: Place *N* queens on the chessboard in such a way that they cannot check each other.

Focus:

1. Different crossover operators – 1-point, 2-point and uniform crossover.

17. Regression Problem

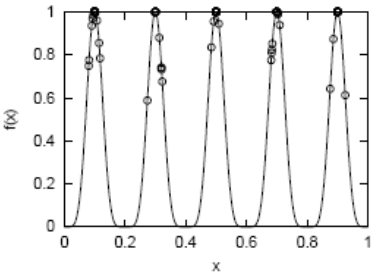
Given: Set of *N* data samples $[x_i,y_i]$.

Goal: Find a function that fits to the training data as best as possible.

Focus:

1. Compare GP evolving the whole function expression with GA evolving only parameters of the polynomial.
2. Try GP with different function sets.

18. Multimodal function Optimization



Given: A multimodal function $F(x)$ where $x \in [LB, UB]$.

Goal: Find as many local optima of the function as possible on the interval $[LB,UB]$.

Focus:

1. SGA vs. GA with fitness sharing strategy

19. Longest common subsequence

20. Minimum Energy Broadcast

21. Mastermind

22. Quadratic Assignment Problem

23. Problém batohu

24. Vícekriteriální problém batohu

25. Vícekriteriální TSP