Genetic Programming

CSI 5388
Paper Presentation

Joe Burpee
2005/2/9

Genetic Programming

- Koza 1992 book
- Genetic Algorithm (Mitchell ch. 9):
  - Hypothesis representation: parse tree
    - Functions
    - Terminals
  - Operators:
    - Crossover: exchange subtrees
    - Mutation: replace subtree randomly
  - Fitness functions

Papers

1. Bijan KHosraviani, Raymond E. Levitt and John R. Koza
Organization Design Optimization Using Genetic Programming
Late Breaking Papers at the 2004 Genetic and Evolutionary Computation Conference, 26 July 2004.

2. Jeroen Eggermont, Joost N. Kok and Walter A. Kosters
Genetic Programming for data classification: partitioning the search space
Proceedings of the 2004 ACM symposium on Applied computing, March 2004.

1. Organization Design using GP

- GP as postprocessor optimizer for project organization design simulator Virtual Design Team VDT (Kunz, CACM, 1998)

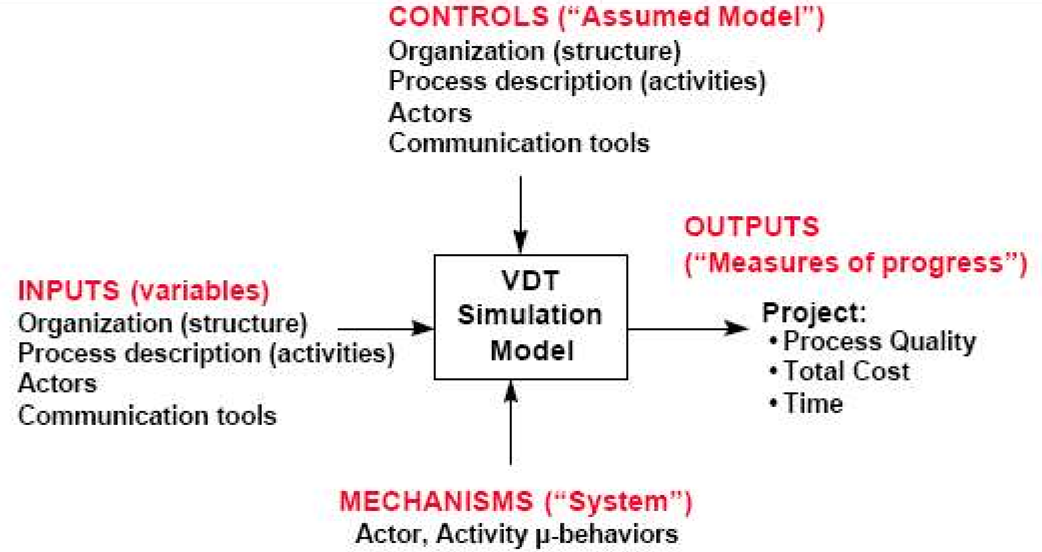

Figure 3. VDT Model Architecture. Given values for independent input variables that describe a project and a set of fixed assumptions, the VDT model simulates each activity being performed by responsible actors and computes overall project duration, cost and coordination quality. The microbehaviors consider both planned direct work and inferred requirement for coordination and rework.

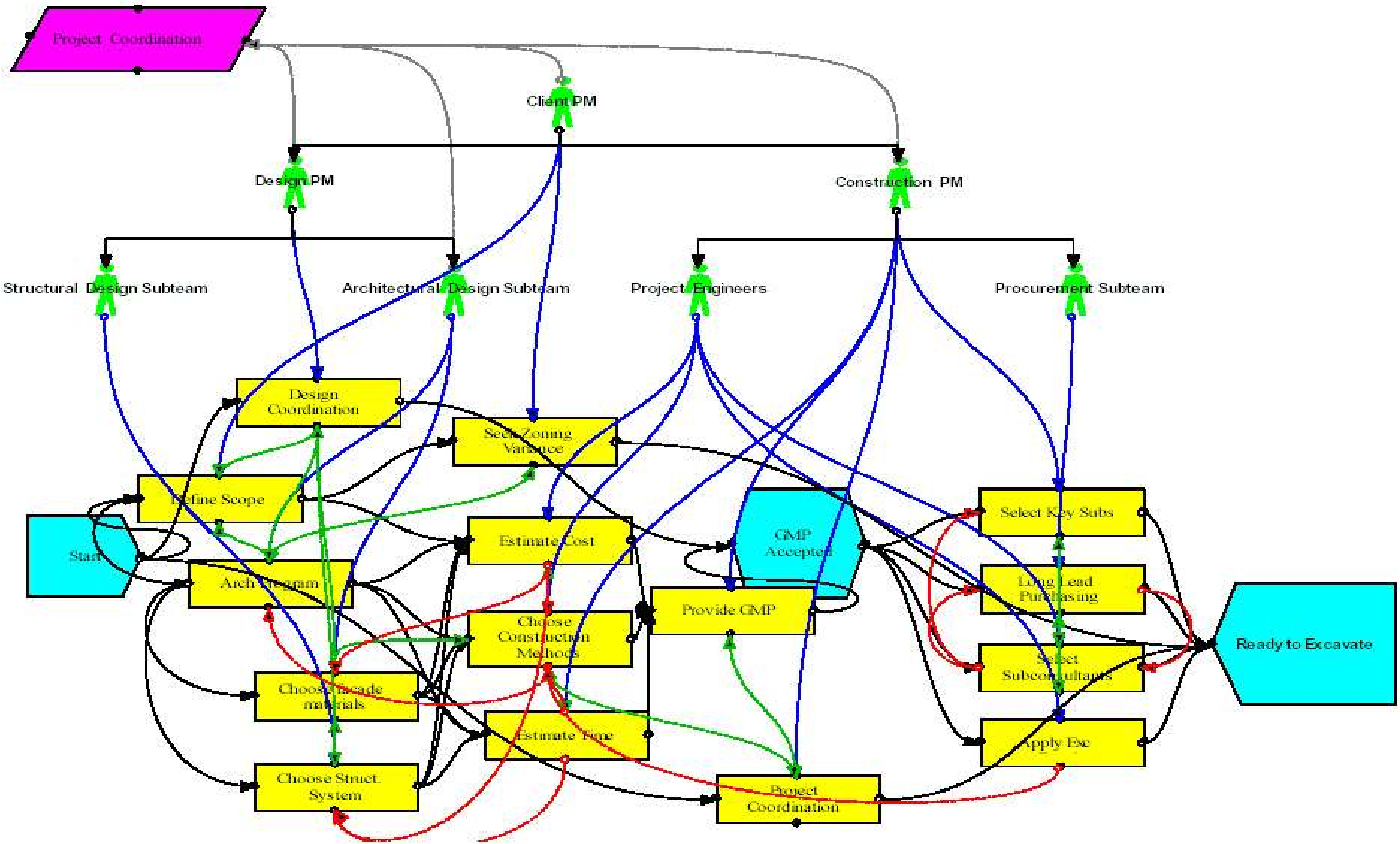

Fig. 1. User Interface of the VDT Simulator - Each project participant fills a position in the project organizational hierarchy and works on one or more activities. The organizational structure and the interdependence between activities define coordination requirements among individuals

VDT Inputs

- Resource counts, budget
- Topology
- Skill levels
- Decision making policies:
  - *Centralization* (delegation)
  - *Formalization* (meetings)
  - *Matrix strength* (collocation)

Fitness =

$$SPD + TFTE * FTEW + \sum_{i=1}^{M} (FRI_i * FRIW_i + PRI_i * PRIW_i + CR_i * CRW_i)$$

SPD = Simulated Project Duration

TFTE = the Total FTE added

FTEW = FTE Weight (TFTE > 3.0 => 1000, else 1)

FRI_i = Functional Risk Index for activity i

$FRIW_i$ = FRI weight for activity i (FRI_i > 0.5 => 1000, else 1)

PRI_i = Project Risk Index for activity i

$PRIW_i$ = PRI weight for activity i (PRI_i > 0.5 => 1000, else 1)

CR_i = Communication Risk for activity i

CRW_i = CR weight for activity i (CR_i > 0.5 => 1000, else 1)

M = maximum number of activities

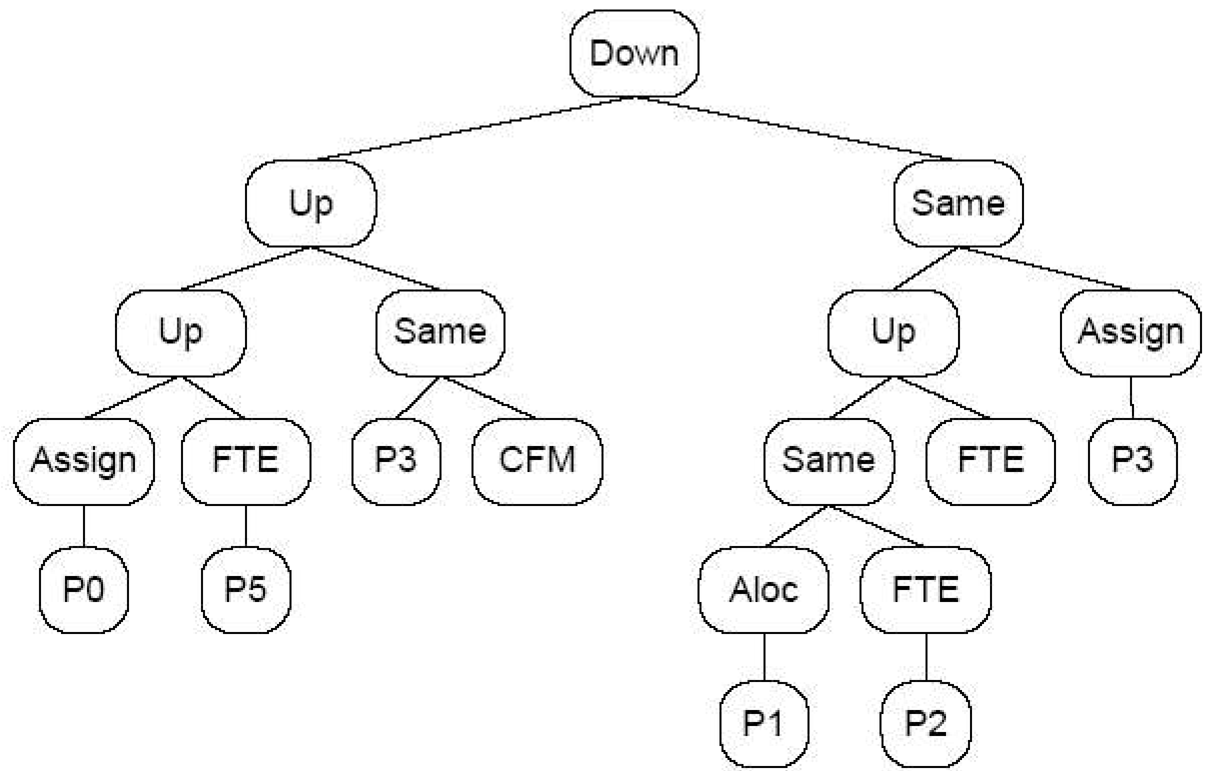

Fig. 3. Sample of a Transforming Genetic Tree. Program trees created by genetic operations modify the structure and attributes of a project organization. The genetic tree above transforms an organization design proposed by a project manager to a near optimal one.

Representation

- Function semantics are obscure:
  - Up, Down, Same have different meaning depending on the Terminals they connect to and whether there are FTE, Assign or Aloc functions in between.
  - E.g. FTE increases or decreases the number of FTEs for each actor depending on the number of Up/Down functions preceding it in the Tree.
- Constraint: FTE, Assign, and Aloc functions can only appear next to the bottom of the tree

Table 1. Tableau for the project organization design optimization problem

Objective:	Find the changes need to be made to the current project organization in order to reduce the project simulated duration, reduce cost and improve quality of the final outcome
Terminal Set	P1, P2, P3, P4, P5, P6, P7, CFM
Function Set	Up, Down, Same, FTE, Assign, Aloc
Fitness Cases	15 total – 1 for simulation duration, 1 for FTE, 13 for each activities
Raw Fitness	SPD + TFTE * FTEW + $\sum$ (FRI_i * $FRIW_i$ + PRI_i * $PRIW_i$ + CR_i * CRW_i) (see section 4.4 Fitness Evaluation)
Standardized Fitness	Same as raw fitness
Parameters	Population size M = 3000 Maximum number of generations, G = 100 Crossover = 90% Mutation = 3% Reproduction = 7%
Success Predict	None – search for the shortest simulation duration with the given quality and FTE constraints

Results

- Experiment 1: varied only skill levels to benchmark known optimum (infinite budget)
- Experiment 2: varied FTEs, policies, to compare with 40+ student/manager team solutions
- Best individual found by GP in generation 21 beats the best human-discovered solution by 2 days (over approx. 8 months)
- Best tree has 99 non-leaf nodes
- Generalization?

The best individual found by GP in generation 21, and it is shown below in a lisp-type format:

```
(Up (Down (Same (Same P5 P4) (Down (Down P1
P5) (Up (FTE P0) (Up (Down (Up (FTE P0) (Down
P5 P5)) (Up (FTE P1) (Up (FTE P0) (Same P3
P6)))) (FTE P5)))))) (Up (Same (Same (Down (Up
(Up (Assign P0) (FTE P1)) (Same (Up (Same
(Down (FTE P4) (FTE P0)) (Down (FTE P2) (Up
(Up P6 (Up (Up P0 (FTE P1)) (FTE P4))) (FTE
P1)))) (Up (FTE P4) (Assign P4))) (Up (Up (Up
(FTE P5) (FTE P5)) (FTE P4)) (Up (FTE P0) (Up
(Assign P0) (Same P5 P4))))))) (Up (FTE P5)
(Aloc P0))) P2) (FTE P0)) (Same (Same (Down
(Up (Up (Assign P0) (Same P5 P4)) (Same (Up
(Same (Up (Assign P0) (Up (Assign P1) (Assign
P0))) (Aloc P1)) (Up (FTE P4) (Assign P4)))
(Up (Up (Up (FTE P5) (FTE P5)) (FTE P4)) (Up
(FTE P0) (Up (Assign P0) (Same P5 P4))))))) (Up
(FTE P5) (Aloc P0))) P2) (FTE P0)))) (FTE P4))
```

2. Decision Trees using GP

Multi-layered Fitness

- Primary: misclassification percentage
- Secondary: number of tree nodes (also pruning)

If individuals tied on primary fitness, compare secondaries.

Full atomic representations

Each node is *attribute operator value*

- Non-leaf (boolean):
  numeric: *attribute* < *value*
  nominal: *attribute* = *value*
- Leaf (assignment): *class* := *C*

Simple Representation

• Potential atoms for every *attribute-value* combination.

• Flexible, but huge search space.

Refined representation

• Analogous to C4.5 but not greedy; uses *gain* or *gain-ratio* but globally.

• Instead of splitting numeric attributes at single threshold, splits them into *k* intervals (k-1 thresholds).

• $k \leq 5$ here.

Clustering representation

Partitions each numeric attribute globally using *k-means*.

Modified Atoms

Refined representation

- (attribute $<$ threshold1),
- (attribute $\in$ [threshold1, threshold2)),
- (attribute $\in$ [threshold2, threshold3)), and
- (attribute $\geq$ threshold3).

Clustering representation

- (attribute $\in$ [min1, max 1]),
- (attribute $\in$ [min2, max 2]), and
- (attribute $\in$ [min3, max 3]),

Table 1: **Example data set**

A	B	class
1	a	yes
2	b	yes
3	a	no
4	b	no
5	a	yes
6	b	yes

Numeric atoms:

- (A <1), (A < 2), (A < 3), (A < 4), (A < 5), and (A < 6)
- *gain-ratio:* (A < 3), (A ∈ [3, 5)), and (A ≥ 5)
- *k-means:* (A ∈ [1, 2]), (A ∈ [3, 4]), and (A ∈ [5, 6])

Experiments

- Mutation = crossover = 90%
- Population 100, generations ≤ 99
- Tournament selection
- Nodes ≤ 63, pruned automatically
- 10-fold crossvalidation
- Performance = average misclassification rate
- UCI datasets, some with C4.5 results

Table 3: Australian credit data set results.

algorithm	k	average	s.d.	best	worst	rank
clustering GP	2	**13.7**	0.8	12.5	14.8	1
clustering GP	3	14.8	0.7	13.8	16.1	3
clustering GP	4	14.8	0.4	14.3	15.7	4
clustering GP	5	15.2	0.7	13.5	15.8	8
refined GP (*gain*)	2	14.2	0.4	13.5	14.9	2
refined GP (*gain*)	3	15.1	0.8	14.9	16.4	7
refined GP (*gain*)	4	14.9	0.9	13.3	16.5	5
refined GP (*gain*)	5	15.1	0.6	13.9	16.4	6
refined GP (*gain_ratio*)	2	15.7	0.4	14.9	16.4	12
refined GP (*gain_ratio*)	3	15.5	0.1	15.4	15.7	9
refined GP (*gain_ratio*)	4	15.5	0.3	15.1	15.9	10
refined GP (*gain_ratio*)	5	15.6	0.4	15.1	16.1	11
simple GP		22.0	3.0	17.0	25.7	14
C4.5 *		15.9				13
Bagged C4.5		N/A				
Boosted C4.5		N/A				
CEFR-MINER		N/A				
ESIA		19.4	0.1			15

Table 7: Ionosphere data set results

algorithm	k	average	s.d.	best	worst	rank
clustering GP	2	13.1	0.9	11.4	14.2	16
clustering GP	3	10.5	1.2	8.8	13.4	9
clustering GP	4	12.1	1.3	9.4	14.0	14
clustering GP	5	13.3	2.1	10.8	17.4	17
refined GP (*gain*)	2	8.3	1.0	7.1	10.8	5
refined GP (*gain*)	3	10.5	1.1	9.1	12.5	10
refined GP (*gain*)	4	10.8	0.6	9.9	12.0	11
refined GP (*gain*)	5	11.6	1.7	8.8	15.1	13
refined GP (*gain_ratio*)	2	7.7	0.7	6.8	9.1	3
refined GP (*gain_ratio*)	3	8.1	0.8	7.1	9.4	4
refined GP (*gain_ratio*)	4	8.3	0.9	6.5	10.0	5
refined GP (*gain_ratio*)	5	9.1	1.0	7.1	10.2	8
simple GP		12.4	1.8	8.0	14.3	15
C4.5		8.9				7
Bagged C4.5		6.2				2
Boosted C4.5		**5.8**				1
CEFR-MINER		11.4	6.0			12
ESIA		N/A				