Advancements in Genetic Programming for Data Classification

Dr. Hajira Jabeen

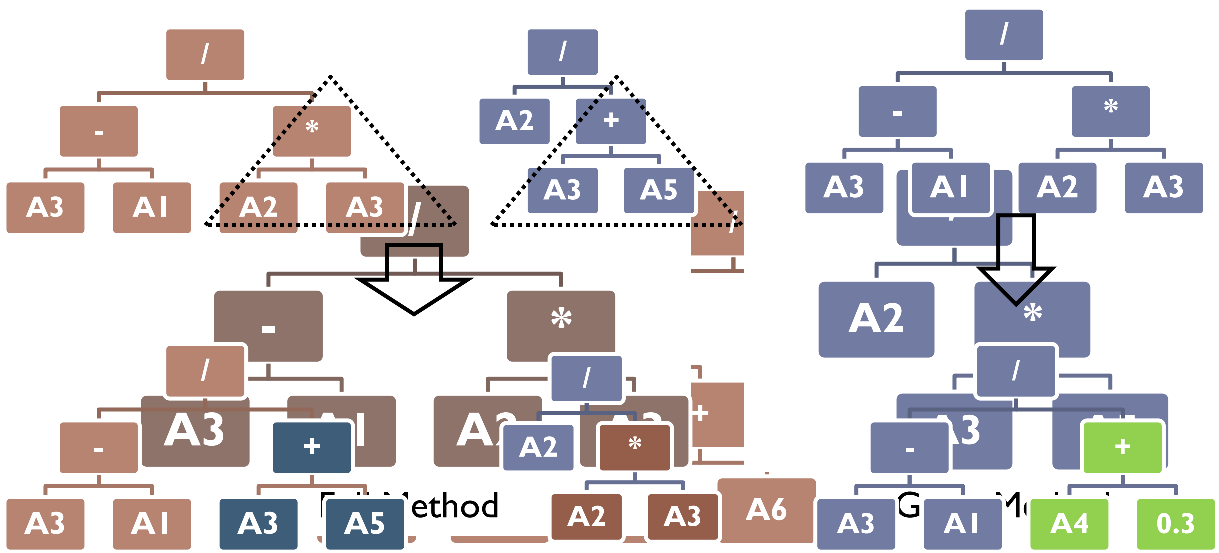
Iqra University

Islamabad, Pakistan

Introduction

Genetic Programming

1. Solution representation
2. Random solution initialization
3. Fitness estimation
4. Reproduction operators
5. Termination condition



Crossover

Mutation

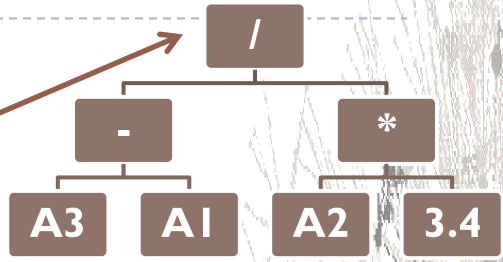
Classification using Genetic Programming

- Advantages
  - Global search
  - Flexible
  - Data modeling
  - Feature extraction
  - Data distribution
  - Transparent
  - Comprehensible
  - Portability
- Disadvantages
  - Long training time
  - Bloat (larger classifier size)
  - Lack of convergence
  - Mixed type data
  - Multiclass classification

DepthLimited Crossover

Classification using GP

- Solution Representation
  - Arithmetic classifier expression (ACE)
  - Function set = {+ , - , / , * }
  - Terminal set = {attributes of data, ephemeral constant}



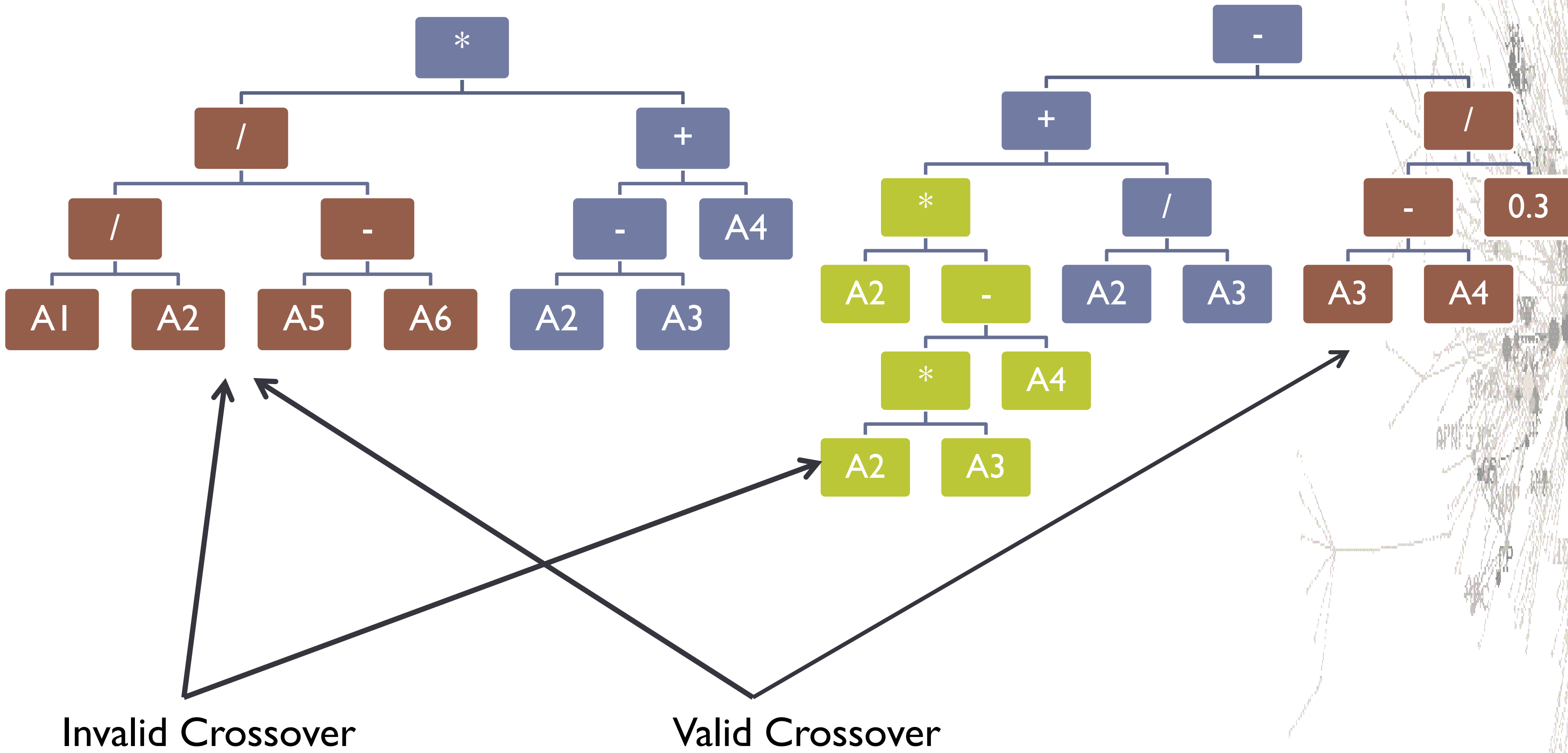
- Solution Initialization
  - Ramped half and half method
- Maximum Depth
  - depth=ceil($\log_2(A_d)$)
  - Where A_d is total number of attributes present in the data

Classification using GP

- Fitness
  - Classification accuracy

- Reproduction operators
  - Mutation
    - Point mutation
  - Reproduction
    - Copy operator
  - Crossover
    - DepthLimited Crossover

DepthLimited Crossover



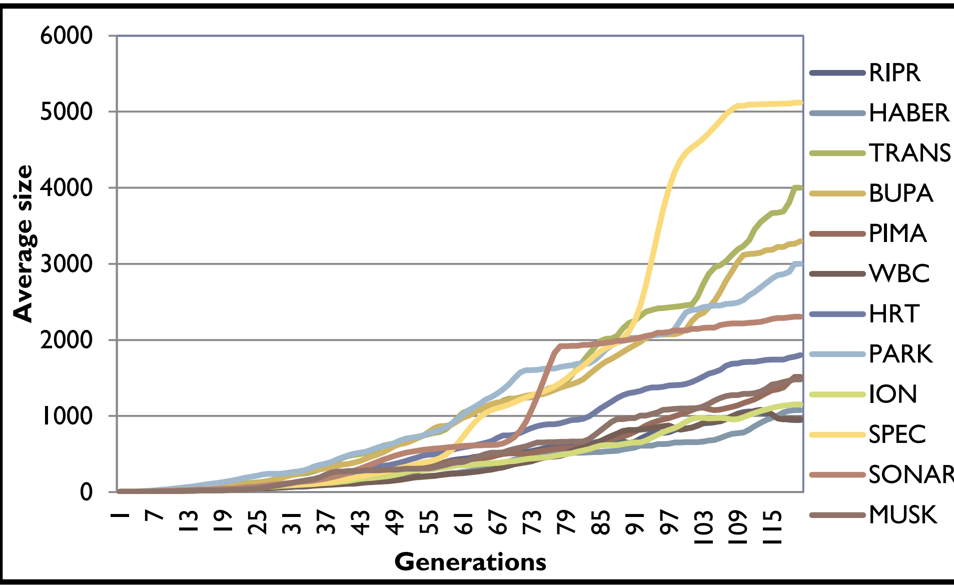
Jabeen, H and Baig, A R., "DepthLimited Crossover in Genetic Programming for Classifier Evolution.“ *Computers in Human Behaviour, Vol (27) 1, pp 1475-1481, Springer, 2010.*

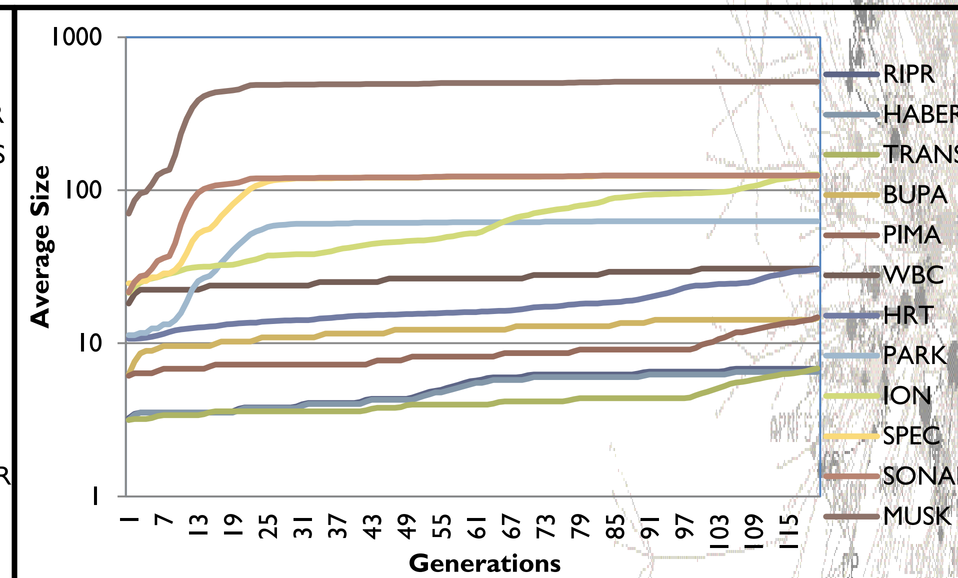
DepthLimited Crossover

- Initial Depth Limits
  - Limit the search space
  - Liberal search limits
  - Depth must include all the attributes
  - The value of depth
  - $d = \mathrm{ceil}(\log_2(A_d))$

  - A full tree of depth ‘d’ can contain all attributes of data.

Complexity





Increase in average number of nodes in population using GP with no size limits

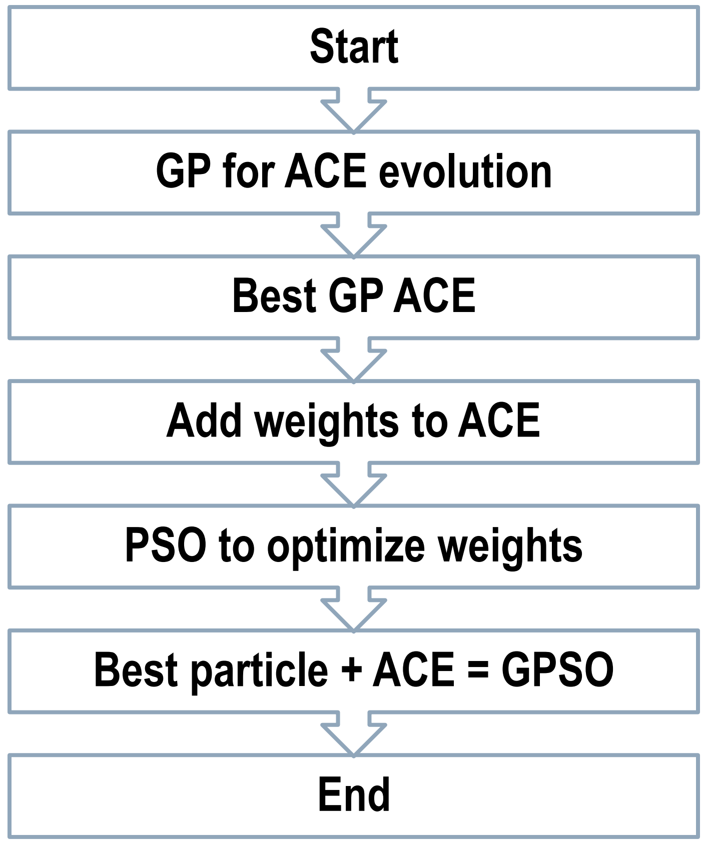
Increase in average number of nodes in population using DepthLimited GP

Optimization of GP Evolved Classifier

Optimization of Expressions

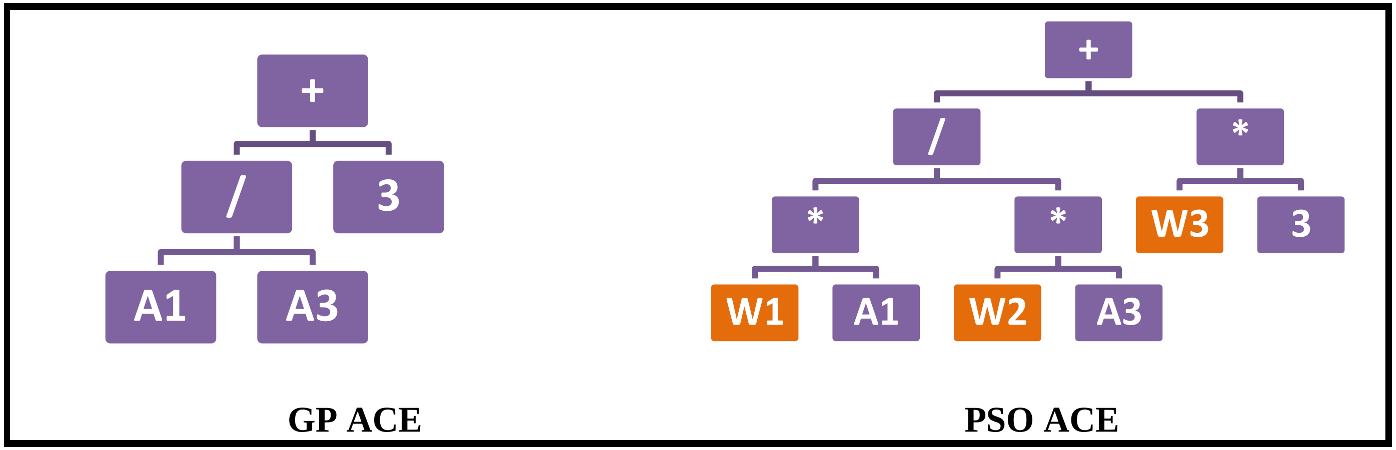
- Motivation
  - Long training time
  - Lack of convergence
- Reason
  - Classifier evaluated for each training instance
  - All evolutionary algorithms do not converge to same solution
- Solution
  - Optimization

GP+PSO=GPSO



Jabeen, H and Baig, A. R., "GPSO: Optimization of Genetic Programming Classifier Expressions for Binary Classification using Particle Swarm Optimization." *International Journal of Innovative Computing, Information and Control, Vol (8) 1a, pp 223-242 2011.*

GP+PSO=GPSO

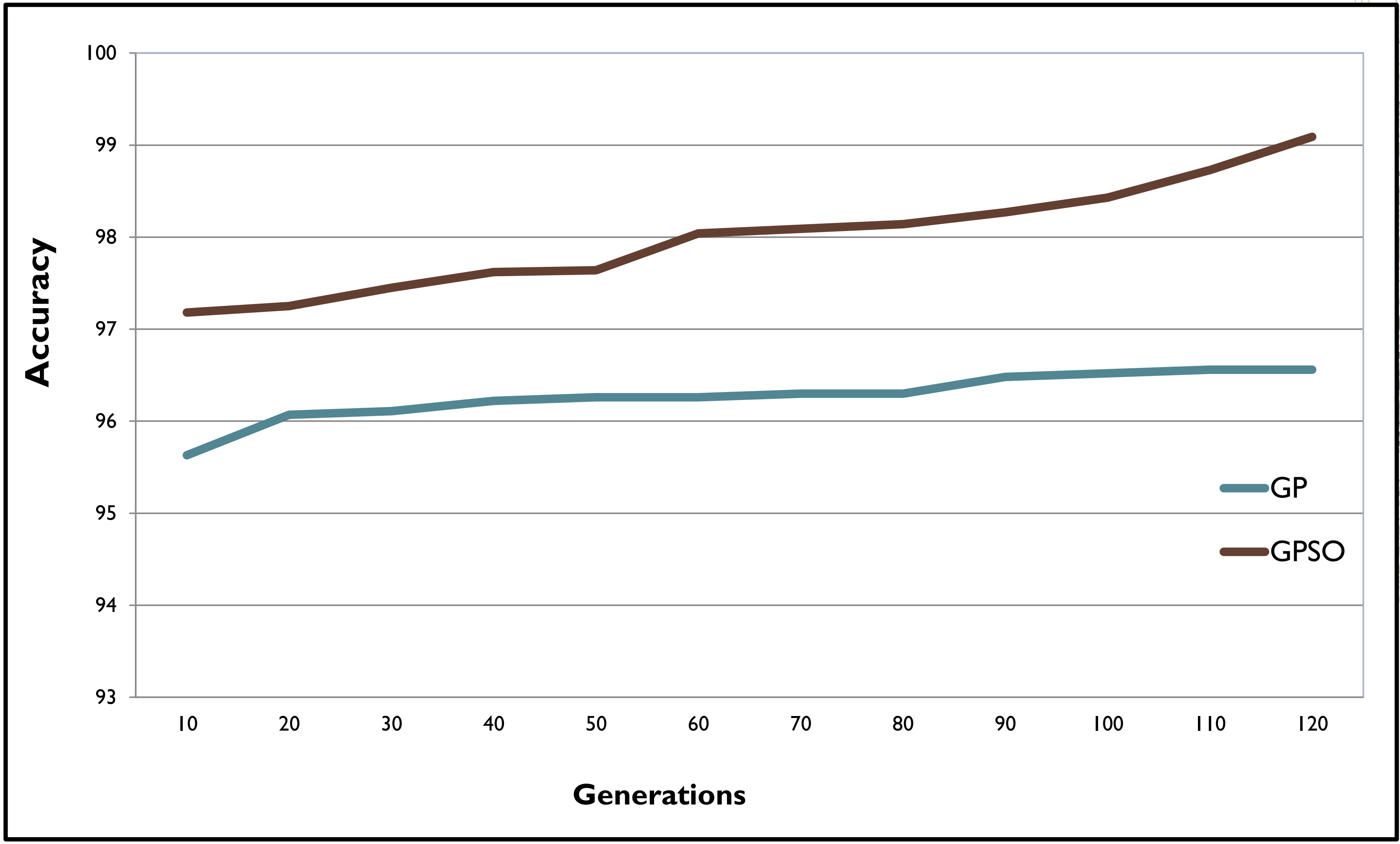


- Solution representation

  = {W1,W2,W3}

- Solution initialization
- Fitness estimation
- Position update
- Termination condition

Wisconsin Breast Cancer Dataset



Classifiers for Pima Indian's Dataset

GP	* + / A3 A1 * A2 A5 + - A3 A2 - A3 8	69.9%
GPSO	* + / * A3 0.50 * A1 0.58 * * A2 0.75 * A5 0.56 + - * A3 0.29 * A2 0.49 - * A3 0.62 * 8 -0.79	71.8%

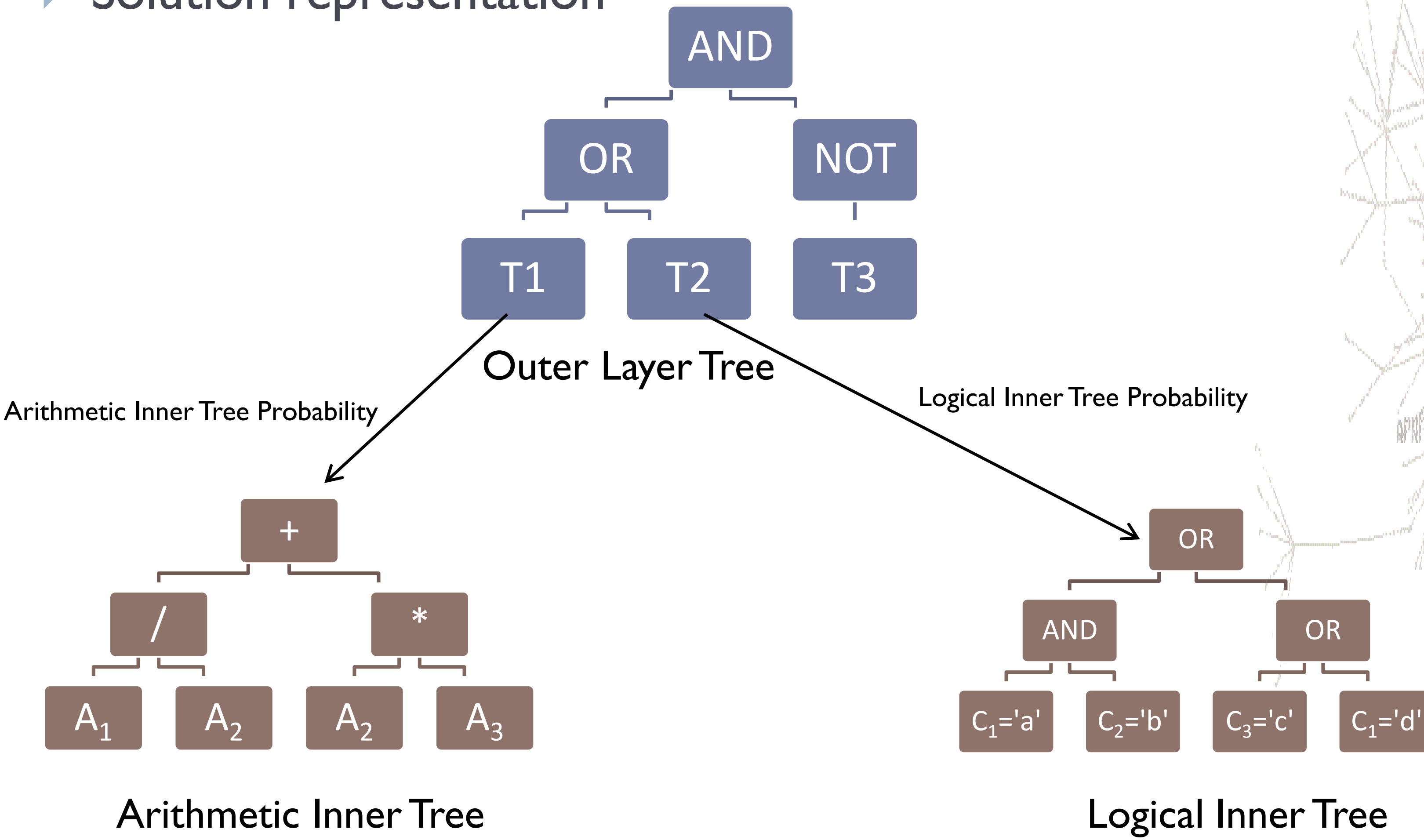
Classification of Mixed-Type Attribute Data

Mixed Attribute Data

- Motivation
  - Use combination of arithmetic expressions and logical rules
- Reasons
  - Classifier expressions applicable to numerical data
  - Rules applicable to categorical data
- Solutions
  - Convert categorical to enumerated value
  - Numerical values to categorical values
  - Use constrained syntax GP

Two Layered Classifier

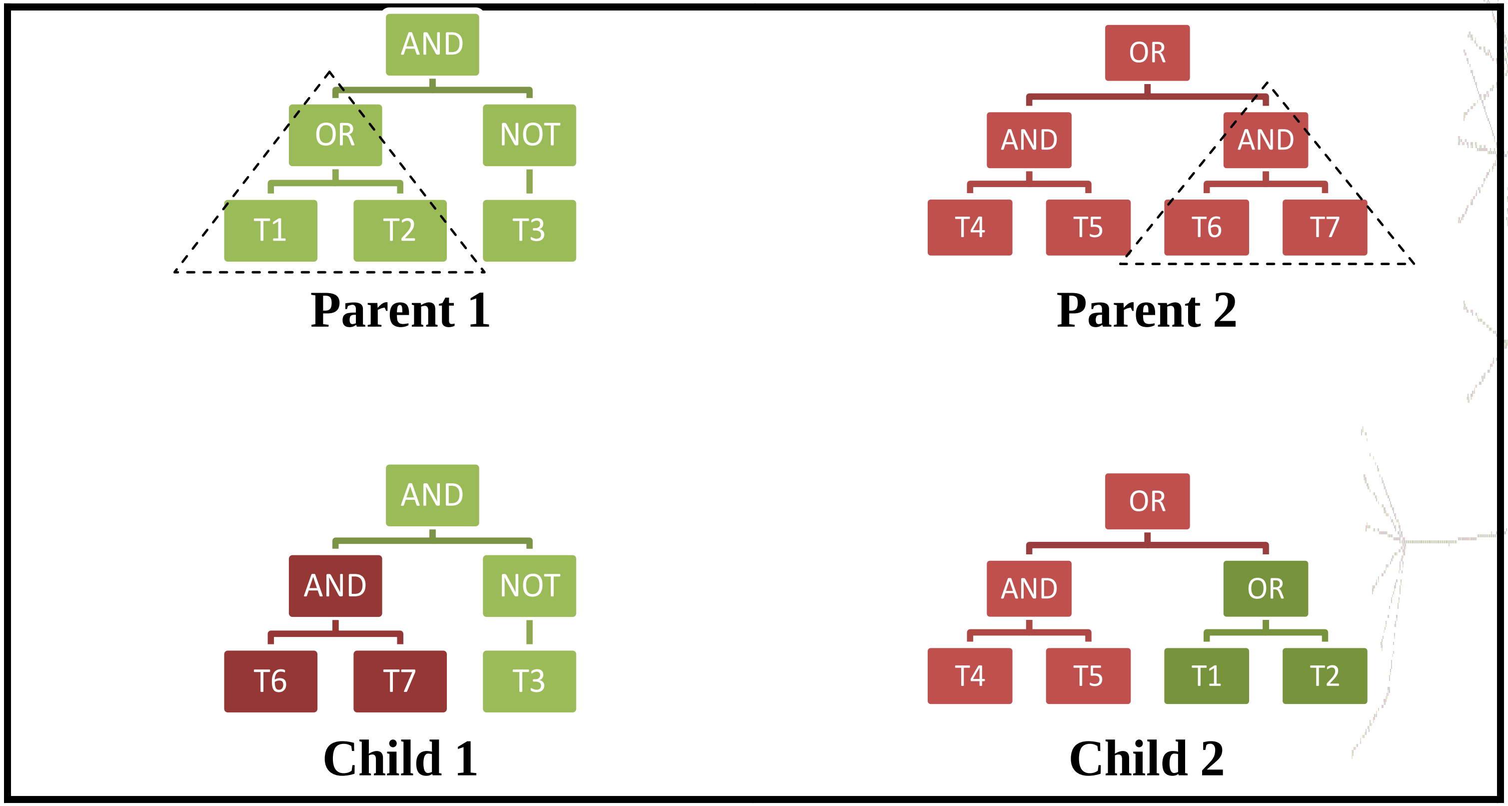
- Solution representation



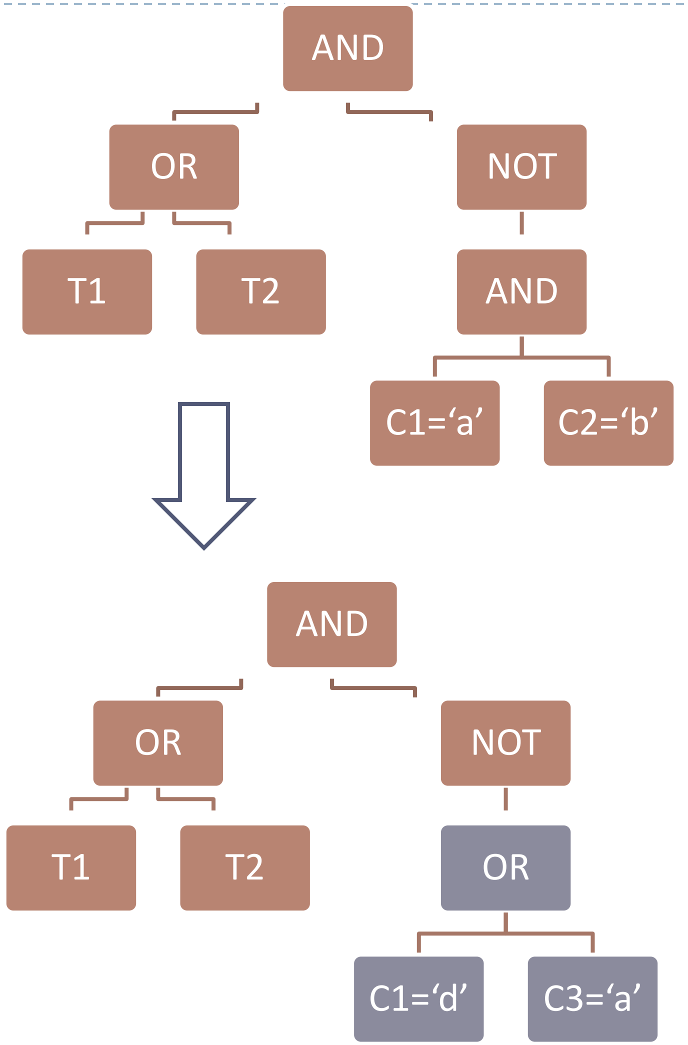
Jabeen, H and Baig, A. R., "Two Layered Genetic Programming for Mixed Variable Data Classification." *Applied Soft Computing, Vol (12) 1, pp 416-422*

Two Layered Classifier

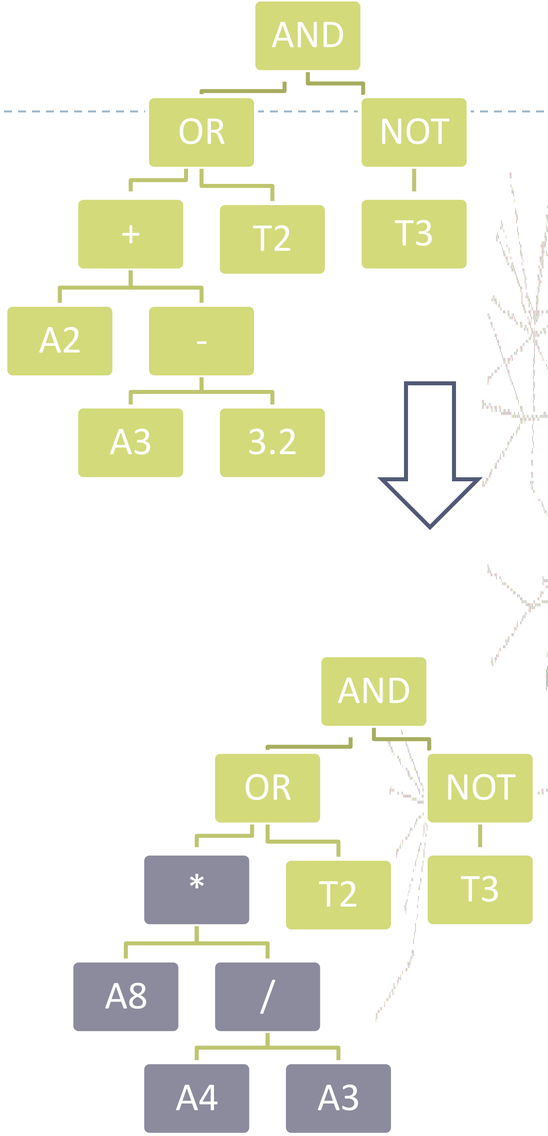
- Crossover Operator



Mutation

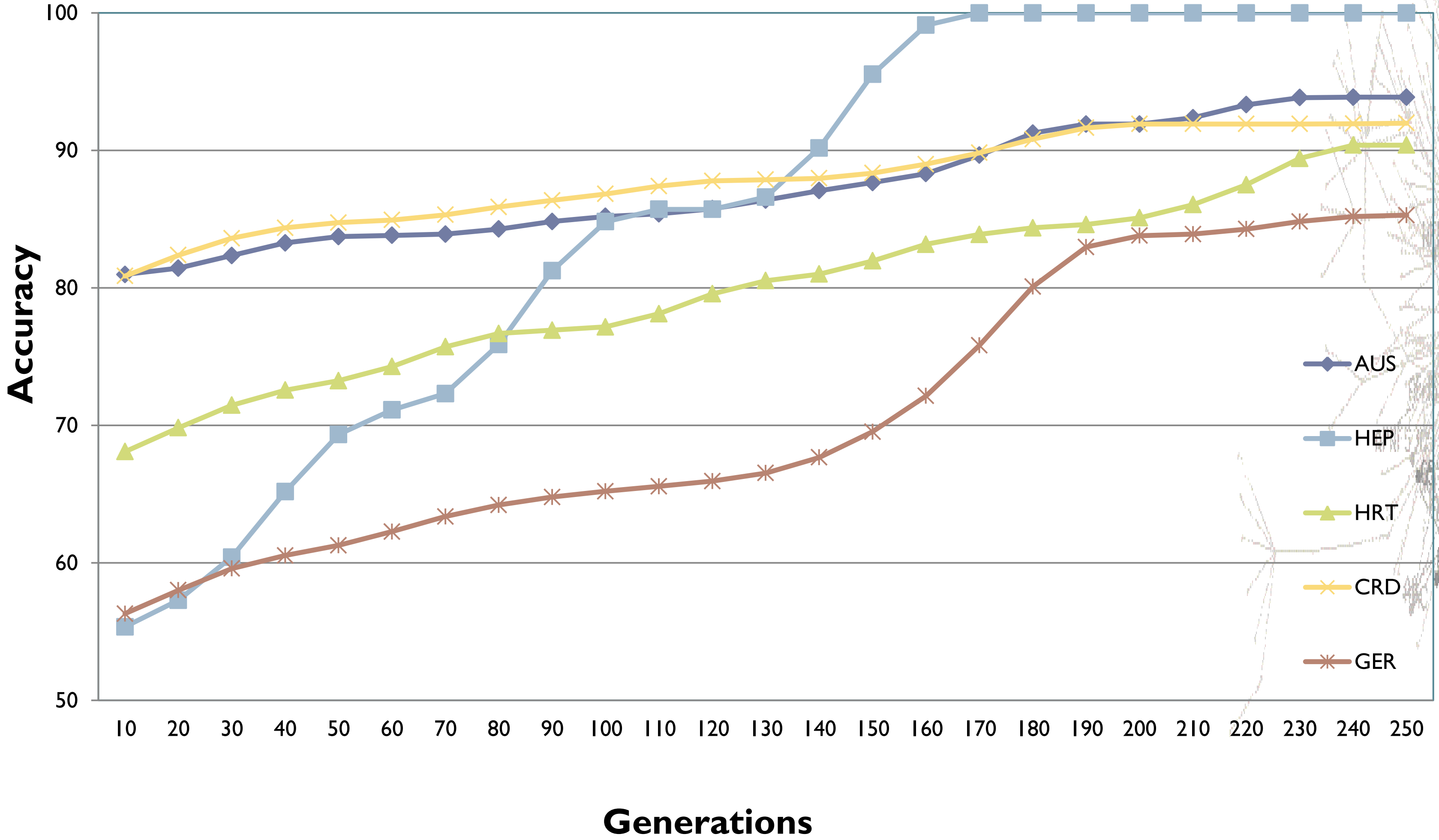


AND
OR
NOT
T1
T2
AND
C1='a'
C2='b'
AND
OR
NOT
T1
T2
OR
C1='d'
C3='a'



AND
OR
NOT
+
T2
T3
A2
-
A3
3.2
AND
OR
NOT
*
T2
T3
A8
/
A4
A3

Evolution of Best Trees



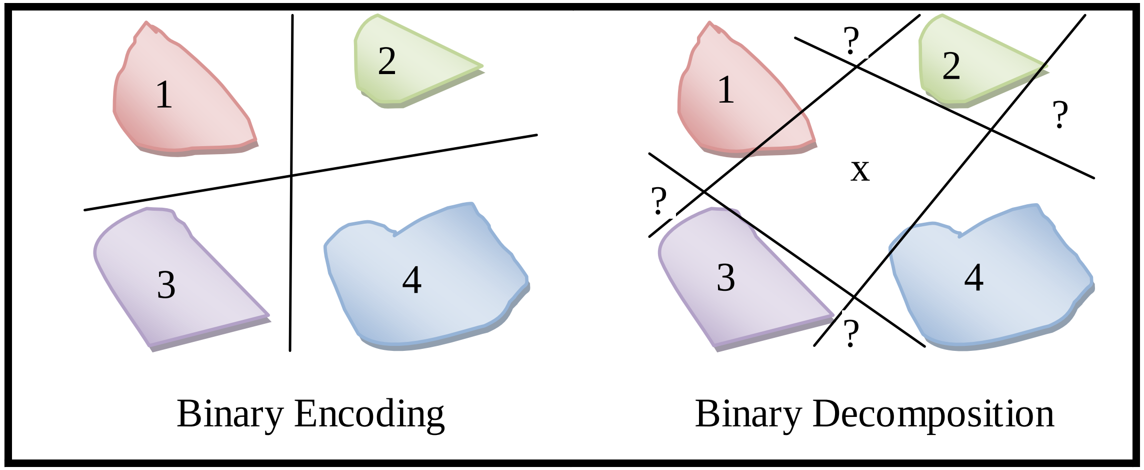
Multi-class Classification

Multi-class Classification

- Motivation
  - Lack of a definitive solution for multiclass classification
- Reasons
  - Binary classifier
- Solutions
  - Binary decomposition
  - Thresholds

Multi-class Classification Approaches

- A four class classification problem, C=4
  - Number of classifiers C in binary decomposition =‘4’
  - Number of classifiers N in binary encoding=ceil($\log_2(4)$)=‘2’
  - Number of conflicts in binary decomposition =12
  - Number of conflicts in binary encoding = 2^N-C=0



Binary Encoding for Classifier

- Classification matrix for a 4 class problem

	Classifier 1	Classifier 2
Class1	0	0
Class2	0	1
Class3	1	0
Class4	1	1

Binary Encoding

	Classifier 1	Classifier 2	Classifier 3	Classifier 4
Class1	1	0	0	0
Class2	0	1	0	0
Class3	0	0	1	0
Class4	0	0	0	1

Binary Decomposition

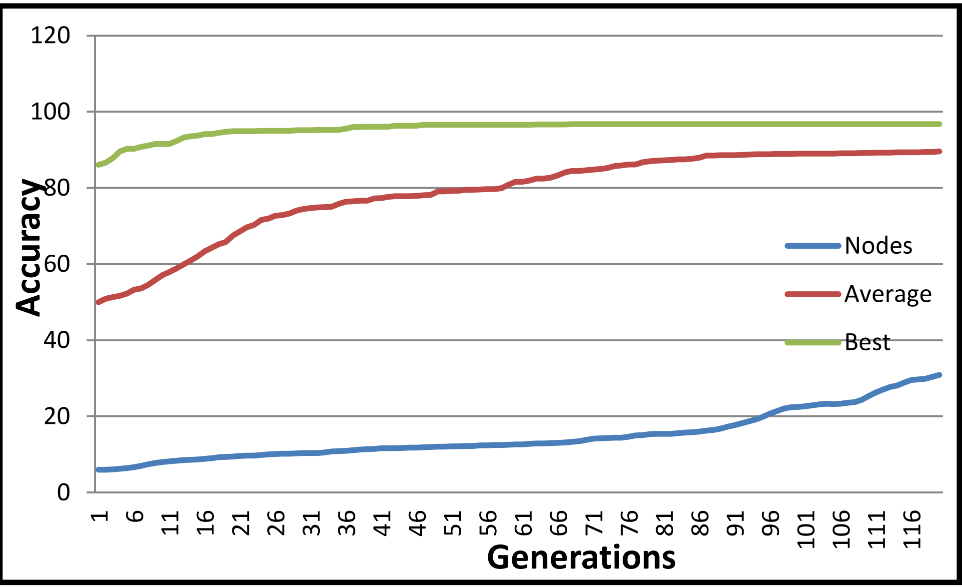
Comparison

BDGP=Binary Decomposition

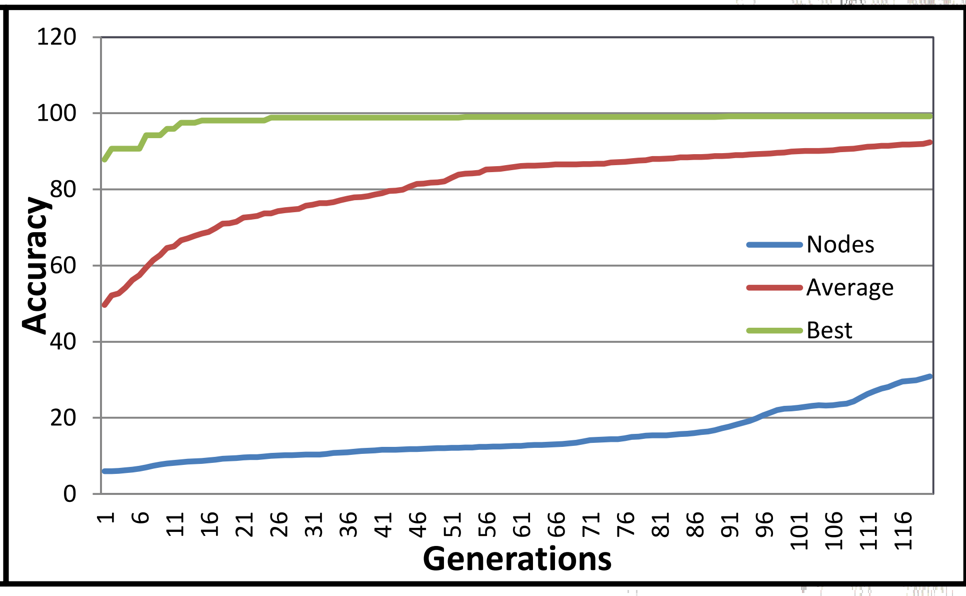
ENGP=Binary Encoding

Datasets	Algorithms	BDGP	ENGP
IRIS	ACC	96.00%	97.20%
	SD	0.2	0.3
	Conflicts	5	1
	Classifiers	3	2
WINE	ACC	78.50%	90.10%
	SD	0.2	0.2
	Conflicts	5	1
	Classifiers	3	2
VEHICLE	ACC	49.00%	83.00%
	SD	0.3	0.2
	Conflicts	12	0
	Classifiers	4	2
GLASS	ACC	54.70%	76.70%
	SD	0.4	0.2
	Conflicts	122	2
	Classifiers	6	3
YEAST	ACC	57.00%	73.80%
	SD	0.6	0.1
	Conflicts	1014	24
	Classifiers	10	4

Complexity and Convergence



Number of nodes and fitness (Average, Best) Wine



Number of nodes and fitness (Average, Best) Iris

Summary

- Computational efficient
- Less conflicts
- Better accuracy

Conclusion

- GP based classification
  - DepthLimited crossover
  - Optimization of classifiers
  - Mixed-type data classification
  - Efficient multi-class classification