Multi-Objective Genetic Programming Projection Pursuit for Exploratory Data Modeling

Ilknur Icke
PhD candidate in Computer Science

The Graduate Center, The City University of New York
365 Fifth Avenue, New York, NY, 10016

Joint work with Dr. Andrew Rosenberg

Women in Machine Learning Workshop, Vancouver 2010

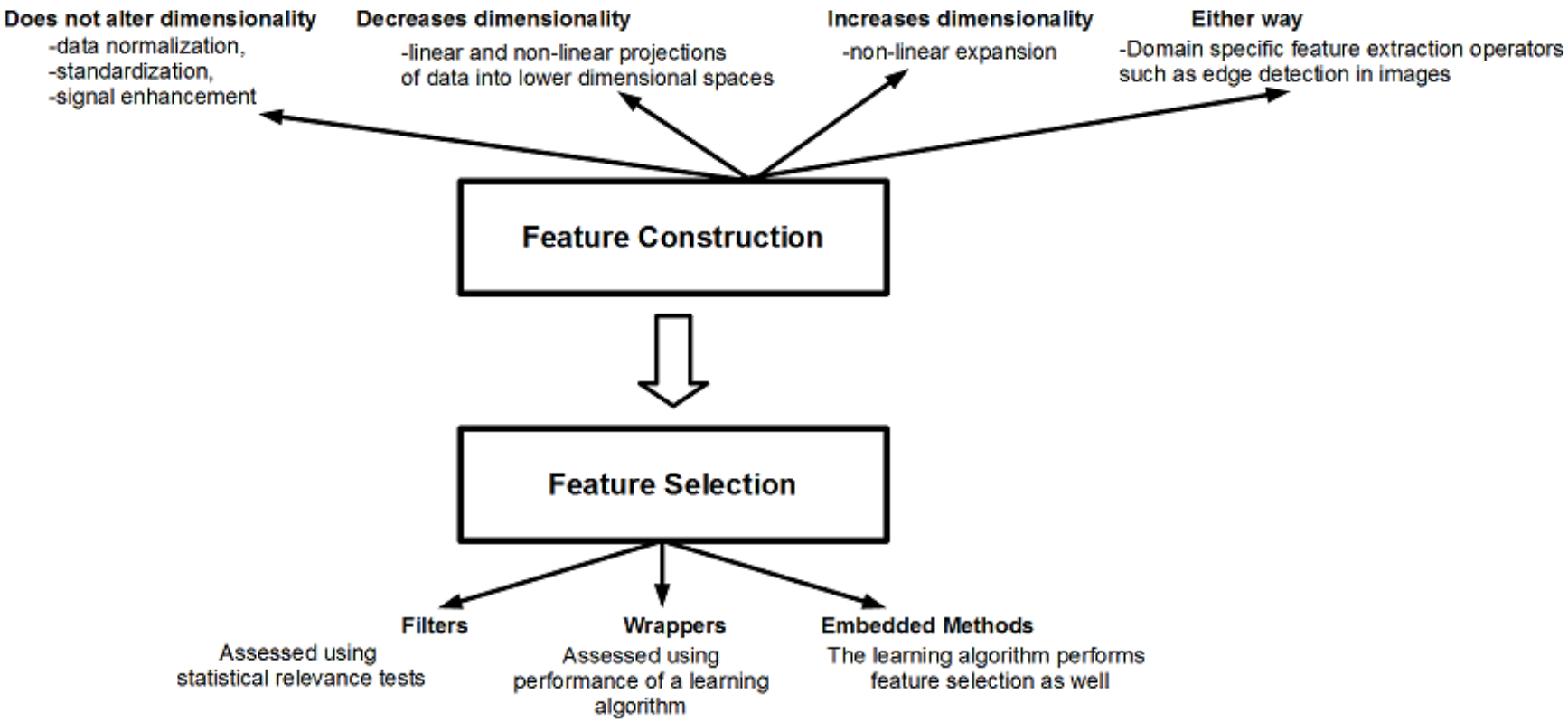
December 6th, 2010

Problem: So much data!

- The *curse of dimensionality* theorem (Bellman,1961): the number of samples needed for a classification task increases exponentially as the number of dimensions (variables, features) increases
- It is costly to collect, store and process data
- Irrelevant and redundant features might hinder performance
- High dimensionality prevents the users from exploring the data visually
- An appropriate feature representation should be selected for each problem

Feature Extraction

Feature extraction is *the process of finding a good representation for a given dataset so that the outcome of the machine learning process is optimized* (Guyon et al.,2006)



The MOG3P algorithm for Exploratory Data Modeling

We use Genetic Programming to evolve transformation functions that project data into 2 or 3 dimensions for visualization

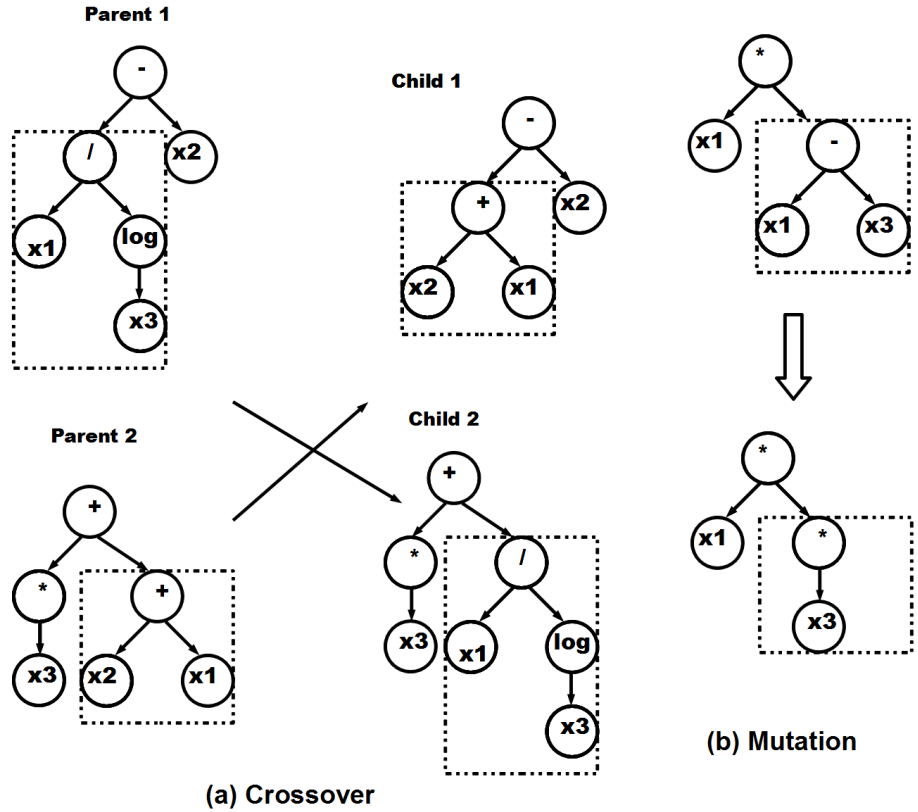
The algorithm (**M**ulti-**O**bjective **G**enetic **P**rogramming **P**rojection **P**ursuit) searches for *interesting* low dimensional projections based on these multiple objectives:

- **classifiability**: the generated data representation should increase the performance of the learning algorithm(s),
- **visual interpretability**: clear class separability when visualized,
- **semantic interpretability** : the relationships between the original and evolved features should be easy to comprehend.

Genetic Programming (GP)

Algorithm 1: The general evolutionary process

Randomly create an initial population of N individuals
foreach *individual in population* **do**
  compute fitness of the individual
end
repeat
  Select individuals from the population as parents
  Perform breeding (crossover,reproduction,mutation operations) to generate new offsprings
  foreach *new individual in population* **do**
    compute fitness of the individual
  end
  Select the best N individuals from the population of parent and offspring individuals as the members of the new generation
until *maximum number of generations are reached or an ideal individual is found*



Crossover and mutation in GP

Multiple objectives and Fitness Function

- **Scalar fitness function**

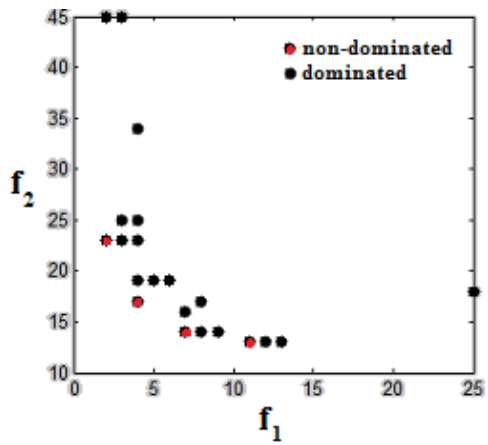
  $F(I) = \alpha * f_1 + \beta * f_2$

  - Advantage: simple fitness comparison
  - Disadvantage: the trade-off (α, β) has to be decided a priori or optimized

- **Vector valued fitness function**

  - Advantage: enables simultaneous optimization of multiple and possibly conflicting criteria
  - Disadvantage: requires more complex comparison schemes such as *pareto dominance*

I_X *dominates* I_Y if $\forall i, I_{X_i} \leq I_{Y_i}$ and $\exists k, I_{X_k} < I_{Y_k}$ where I_{X_i}, I_{Y_i} are the individual criteria. The set of all *non-dominated* solutions are called the pareto-optimal set



Components of the MOG3P Algorithm

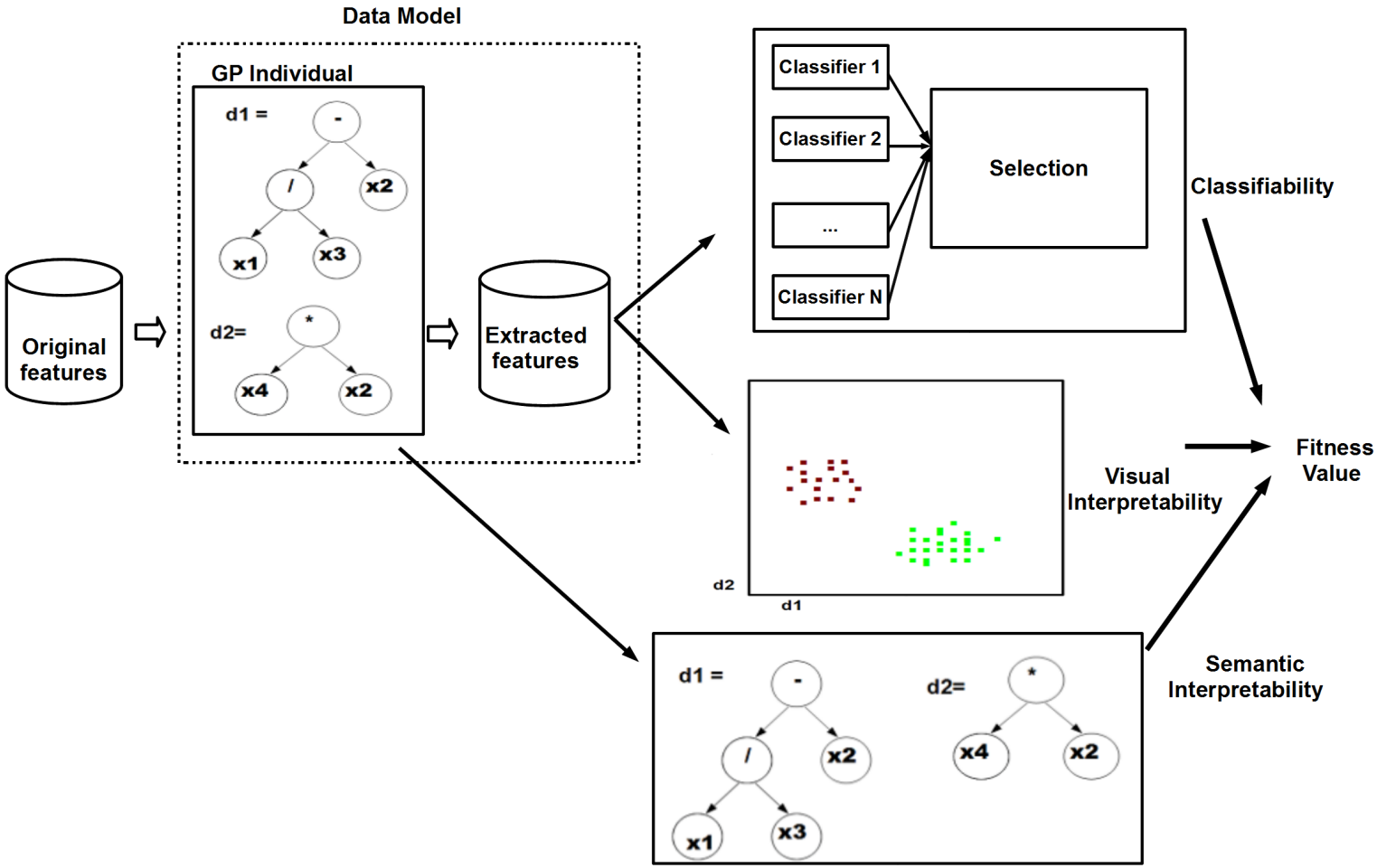


Figure: MOG3P diagram

Some experiments

Table: MOG3P Settings

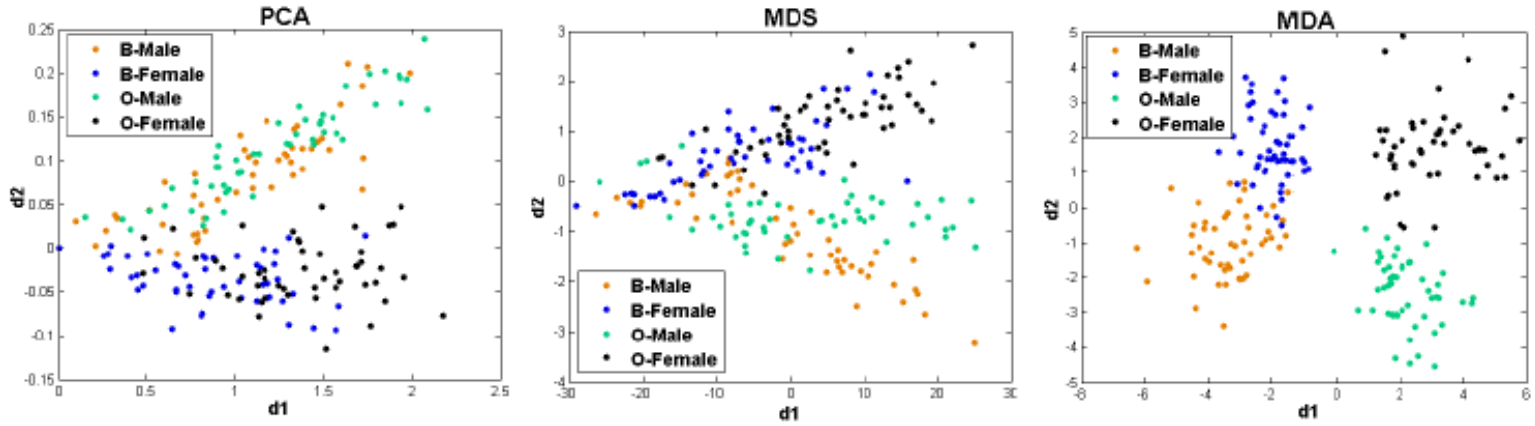
Population Size	400
Generations	100
Crossover Probability	0.9
Reproduction Probability	0.05
Mutation Probability	0.05
Multi objective fitness scheme	SPEA2
Archive Size	100
Basis functions	$\{+, -, *, protected/, min, max, power, log\}$
ERC	$[-1, 1]$
Classifiability Objective (C)	accuracy of a random classifier per individual
Visualization Objective (V)	T=2, $\{ I_{LDA}, I_C, I_{DB}, I_{DUNN} \}$
Semantic Objective (S)	$\{I_{TS} \}$
Cross Validation	10 times 10-fold cross validation (total 100 runs)

$\{ I_{LDA}, I_C, I_{DB}, I_{DUNN} \}$: cluster validity indices measuring compactness/separation

$\{I_{TS} \}$: expression complexity measure

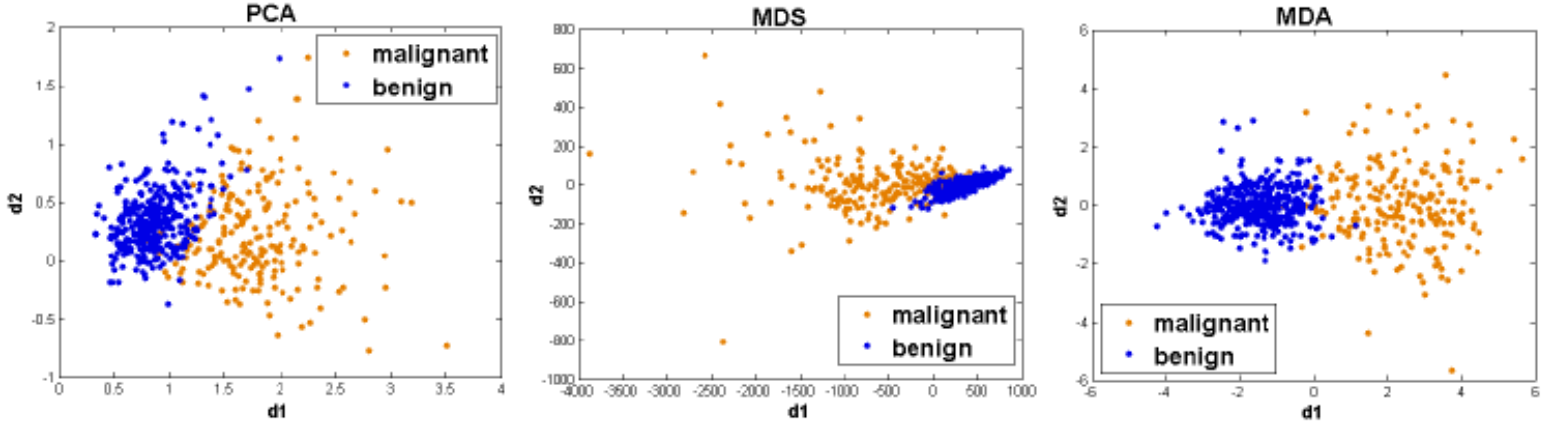
Hypothesis 1: MOG3P can generate better data representations compared to standard methods (PCA, MDS and MDA)

Hypothesis 2: Incorporating an explicit visualization criterion increases performance



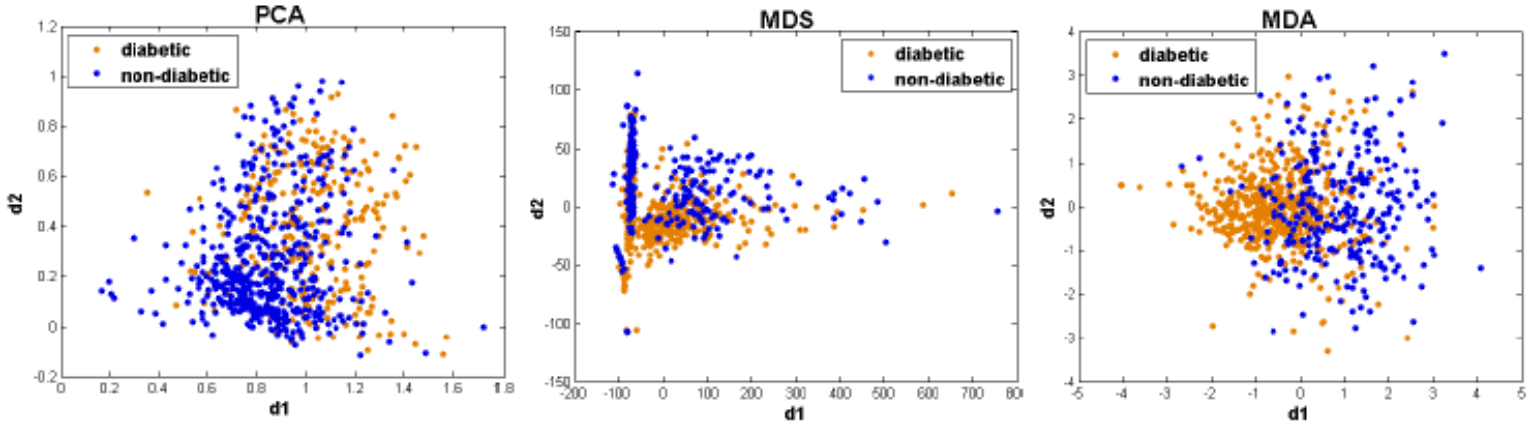
Classifier	PCA (2D)	MDS (2D)	MDA (2D)	All features	MOG3P Random I_{LDA},I_{TS}	MOG3P Random I_{C},I_{TS}	MOG3P Random I_{DB},I_{TS}	MOG3P Random I_{Dunn},I_{TS}
N. Bayes	57.5	67	93.5	38	97.4	96.7	96.25	96.25
Logistic	59.5	63	94.5	96.5	98	96.95	96.6	96.25
SMO	54.5	59	94.5	63.5	97.15	96.45	96.15	95.55
RBF	67	69	96	49	97.65	96.75	96.7	96.25
IBk	57	67.5	93	89.5	97.5	97.35	96.8	96.65
CART	57.5	61	94	75.5	97.05	96.6	96.2	96.1
J48	56.5	59	92.5	73.5	97.35	97	96.25	96.45
Avg	58.5	63.64	*94*	69.36	**97.44**	**96.83**	**96.42**	**96.24**
(std)	(4.03)	(4.19)	*(1.16)*	(20.93)	**(3.62)**	**(4.13)**	**(4.2)**	**(4.47)**

Figure: Crabs Dataset (5-dimensions)



Classifier	PCA (2D)	MDS (2D)	MDA (2D)	All features	MOG3P Random I_{LDA},I_{TS}	MOG3P Random I_{C},I_{TS}	MOG3P Random I_{DB},I_{TS}	MOG3P Random I_{Dunn},I_{TS}
N. Bayes	92.09	90.69	97.19	92.79	97.98	98.21	97.89	97.96
Logistic	94.38	93.15	97.72	94.73	97.89	98.28	98.3	98.37
SMO	93.32	87.52	96.84	98.07	97.61	97.87	97.54	97.66
RBF	94.73	92.79	98.24	93.5	98.21	98.28	98.08	98.21
IBk	91.21	90.86	96.66	94.73	98.17	98.40	98.3	98.3
CART	92.79	91.74	96.66	92.97	97.95	98.03	97.86	98.03
J48	92.97	92.44	97.19	93.67	97.98	98.14	97.89	98.01
Avg	93.07	91.31	*97.21*	94.35	**97.97**	**98.17**	**97.98**	**98.08**
(std)	(1.22)	(1.9)	*(0.6)*	(1.8)	**(2.14)**	**(1.81)**	**(1.83)**	**(1.75)**

Figure: Wisconsin Breast Cancer Diagnostic Dataset (30-dimensions)



Classifier	PCA (2D)	MDS (2D)	MDA (2D)	All features	MOG3P Random I_{LDA},I_{TS}	MOG3P Random I_{C},I_{TS}	MOG3P Random I_{DB},I_{TS}	MOG3P Random I_{Dunn},I_{TS}
N. Bayes	71.22	74.22	76.43	76.30	82.03	81.90	81.68	81.86
Logistic	72.00	74.35	78.26	77.21	81.54	81.39	81.47	81.75
SMO	72.14	74.74	77.73	77.34	81.51	81.25	81.30	81.24
RBF	72.4	73.44	76.82	75.39	82.17	82.41	81.55	81.79
IBk	61.07	64.45	68.36	70.18	82.01	82.44	80.79	81.32
CART	68.49	73.7	75.39	75.13	81.79	81.96	81.15	81.41
J48	69.40	71.88	75.65	73.83	81.58	81.81	80.60	80.93
Avg	69.53	72.4	75.52	75.06	**81.81**	**81.87**	**81.22**	**81.47**
(std)	(4.01)	(3.62)	(3.32)	(2.48)	**(4.32)**	**(4.19)**	**(4.26)**	**(4.32)**

Figure: Pima Indians Diabetes Dataset (8-dimensions)

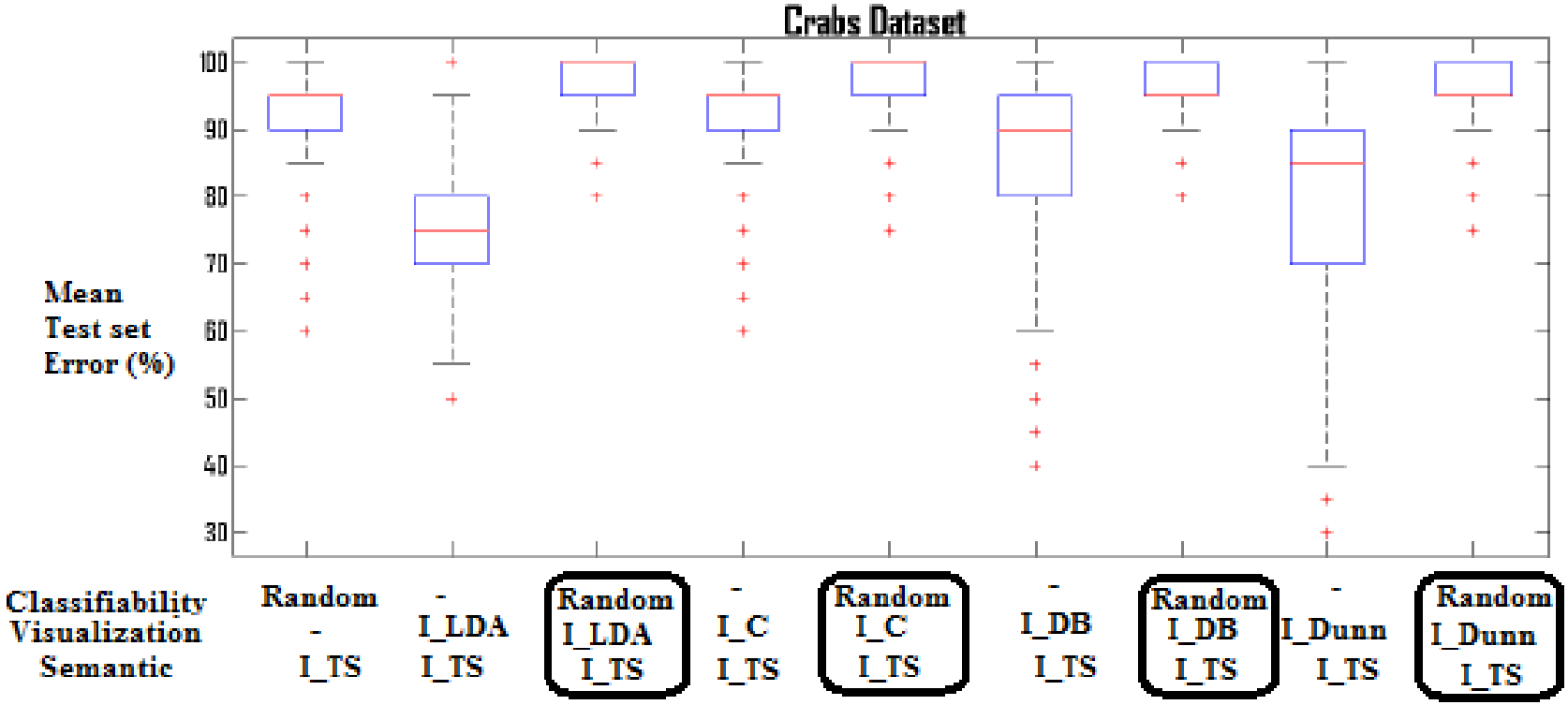


Figure: MOG3P experiments on Crabs Dataset

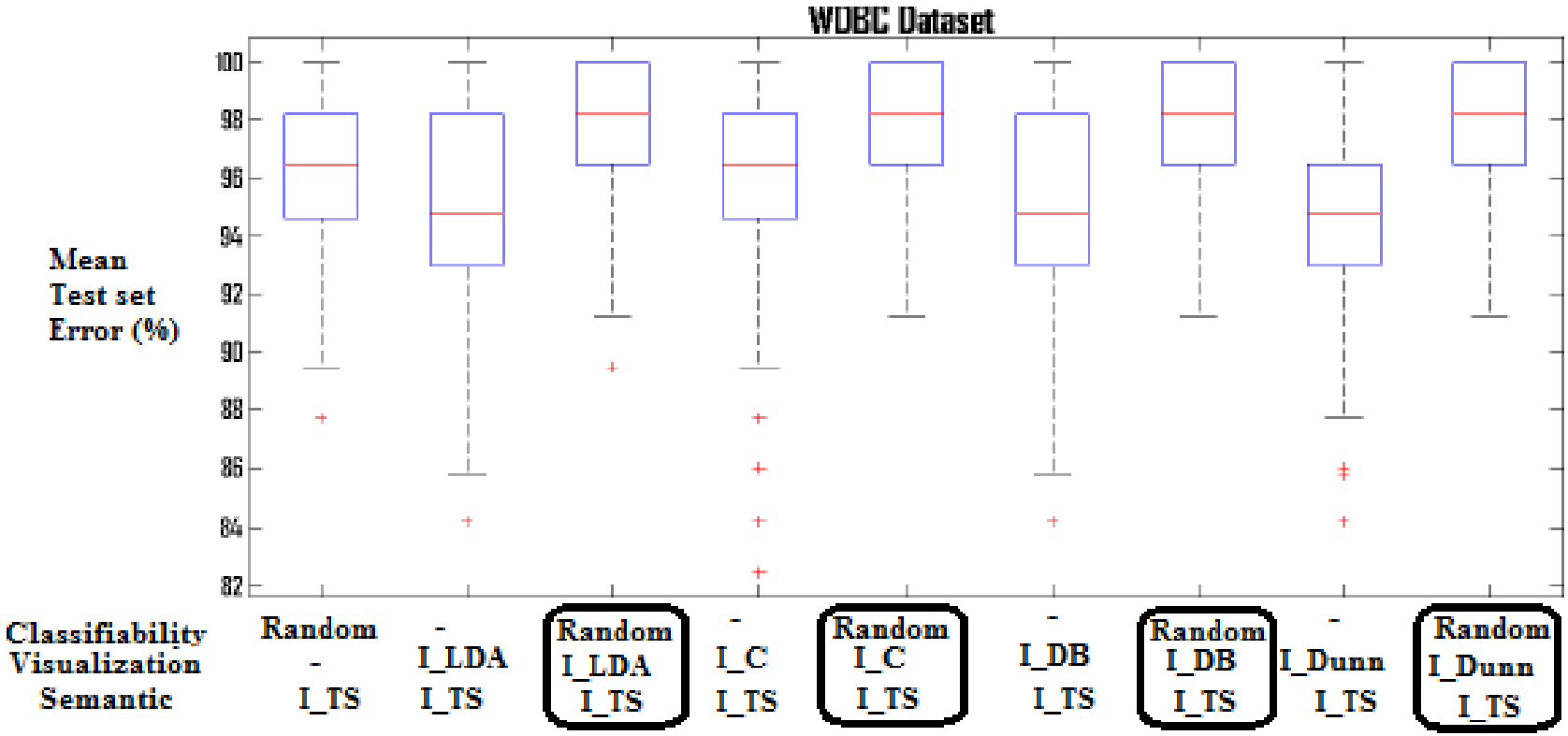


Figure: MOG3P experiments on Wisconsin Breast Cancer Diagnostic Dataset

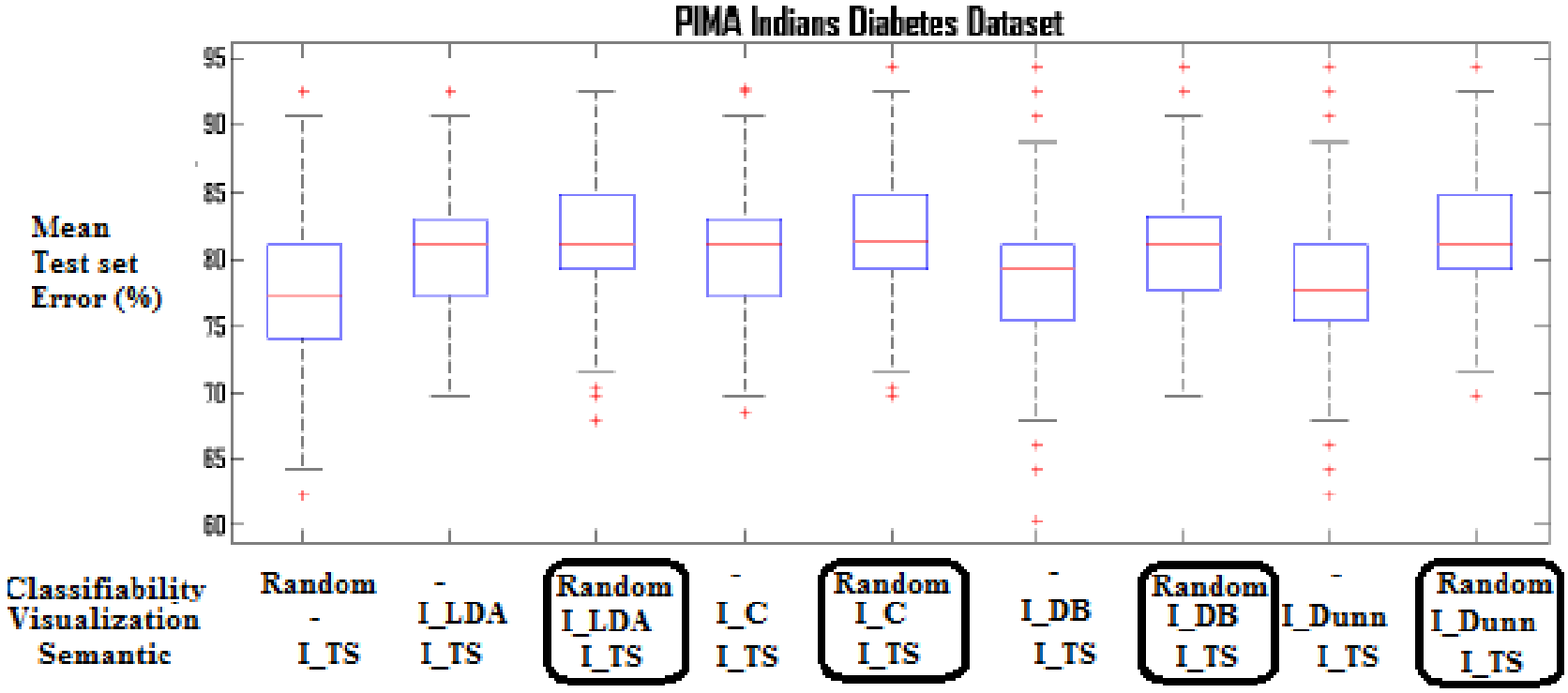


Figure: MOG3P experiments on Pima Indians Diabetes Dataset

Pros & Cons

- Various objectives can be plugged in easily
- Multi-objective
- Model selection options based on the set of pareto optimal candidates:
  - Elimination of unnecessary original features
  - Classifier selection
- Selection of optimal parameters for the problem:
  - individual selection, crossover and mutation operators
  - the fitness function
  - Population size versus generations (small population longer runs vs. larger population shorter runs)
  - Proper building blocks (basis functions)
- High computational cost (fortunately highly parallel)
  - Cluster computing
  - Graphical Processing Units (GPU)

Thank you!

iicke@gc.cuny.edu