Using Genetic Algorithms to Generate Estimation Models

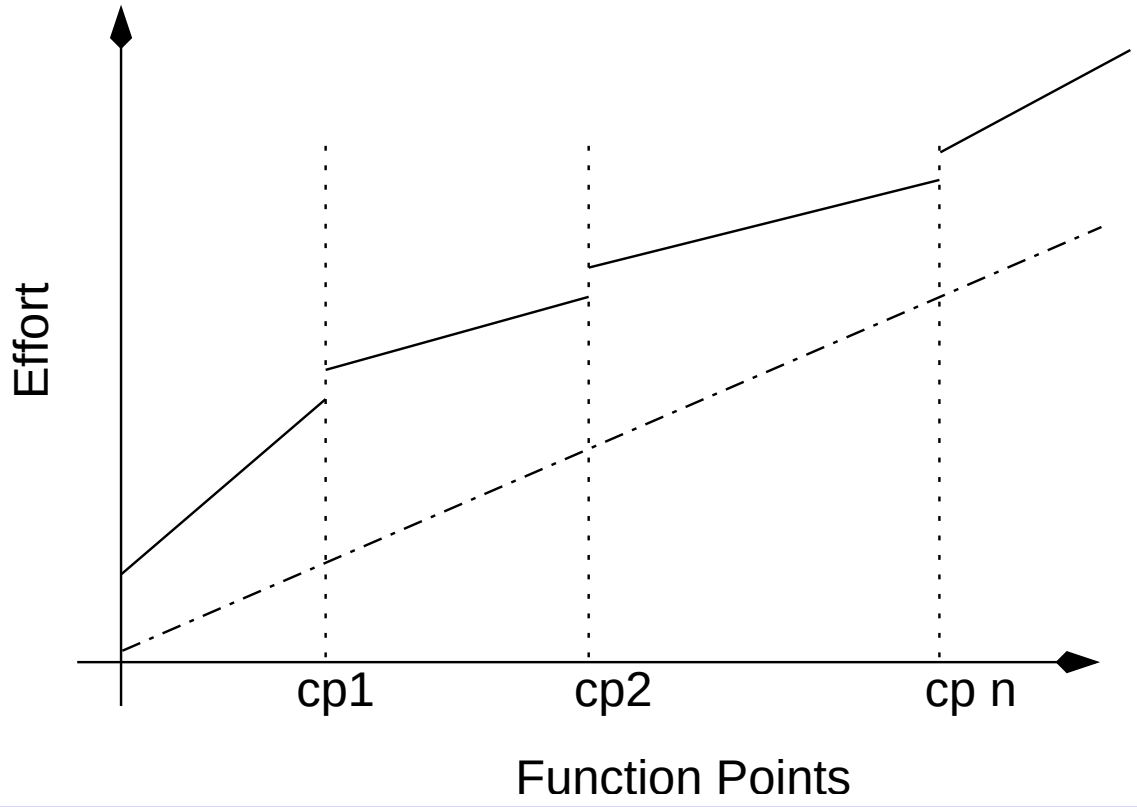
D Rodríguez[1], **JJ Cuadrado-Gallego**[1], J Aguilar[2]
[1]University of Alcala
[2]University Pablo de Olavide

Agenda

- Introduction
- GA in a Nutshell
- The International Software Benchmarking Standards Group (ISBSG)
- GA algorithms to generate multiple estimation models
- Conclusions and Future Work

Introduction

Idea: generate cut-point to divide the dataset using GA to minimize the error



The ISBSG Repository

The ISBSG maintains a software project management repository from a variety of organizations. The ISBSG v8 contains 2028 projects and more than 55 attributes per project classified as:

- Project context such as type of organization, business area, and type of development.
- Product characteristics such as application type user base.
- Development characteristics such as development platform, languages, tools, etc.
- Project size data: different types of function points (IFPUG, COSMIC, etc.)
- Qualitative factors such as experience, use of methodologies, etc.

Datasets. Preprocesing

The first step in all data mining algorithms, it is to preprocess the data. This consists of transforming the data into Weka format selection attributes and instances according to the following criteria:

- Only projects classified as A or B (there are no important inconsistences in the data collection)
- IFPUG – larger dataset including variants NESMA, Albretch or Dreger
- Dependant variable – Normalized Work Effort

Datasets generated from ISBSG

Reality Dataset – DS1 This dataset is provided the ISBSG as part of the *Reality Checker* tool provided as part of the repository. The Reality dataset if composed of 709 instances and 6 attributes (*DevelopmentType*, *DevelopmentPlatform*, *LanguageType*, *ProjectElapsedTime*, *NormalisedWorkEffort*, *UnadjustedFunctionPoints*).

NormWE Dataset– DS2 This dataset is composed of 1390 instances and 15 attributes (*FP*, *VAF*, *MaxTeamSize*, *DevelopmentType*, *DevelopmentPlatform*, *LanguageType*, *DBMSUsed*, *MethodUsed*, *ProjElapTime*, *ProjInactiveTime*, *PackageCustomisation*, *RatioWEProNonPro*, *TotalDefectsDelivered*, *NormWorkEff*, *NormPDR*).

Datasets generated from ISBSG

- Choose initial population
- Repeat
  - Evaluate the individual fitnesses of a certain proportion of the population
  - Select pairs of best-ranking individuals to reproduce
  - Apply crossover operator
  - Apply mutation operator
- Until terminating condition

Parameters GA

- Coding – natural numbers as it is the *'natural'* way for representing FP
- Crossover – uniform mutation max. 30%
- Fitness Function – minimisation of the absolute relative error: $\frac{|p_1-a_1|+...+|p_n-a_n|}{|a_1-\overline{a}|+...+|a_n-\overline{a}|}$

Parameter	**Value**
Population size	20
Generations	50
Crossover probability	0.6
Individual Mutation probability	0.2

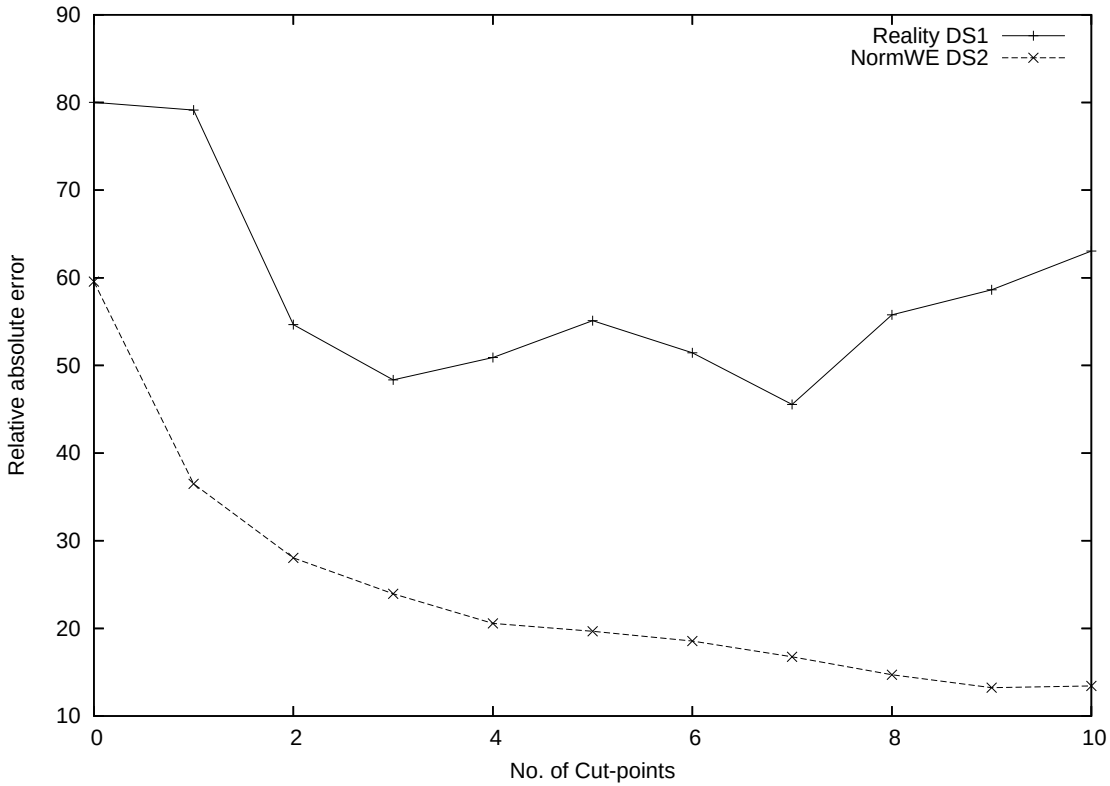
Datasets generated from ISBSG

	DS1	DS2
Correlation coefficient	0.438	0.781
Mean absolute error	5100.2	3420.50
Root mean squared error	11973.86	7463.64
Relative absolute error	80.01%	59.56%
Root relative squared error	89.9%	62.44%

Datasets generated from ISBSG

	Error	Cut-Points in the FP axis for DS2
1	36.48	[2562]
2	28.05	[2081, 2238]
3	23.94	[1235, 1280, 1408]
4	20.57	[1206, 1250, 1373, 1667]
5	19.66	[342, 535, 1039, 2041, 2474]
6	18.55	[339, 963, 1238, 1594, 1781, 2185]
7	16.75	[507, 1260, 1395, 1462, 2047, 2090, 2359]
8	14.71	[356, 688, 1180, 1303, 1655, 1734, 2014, 2091]
9	13.23	[321, 630, 863, 890, 1014, 1881, 2004, 2160, 2663]
10	13.43	[229, 461, 705, 919, 962, 1218, 1389, 1666, 1893, 2484]

Error Summary



Summary and Future Work

- GA can be used to divide the size of dataset to improve estimates
- Theoretically, the error could be reduced as much as needed. In practice, it seems to reach a maximum

Future work includes:

- Use other models other than Linear Regression...
- ...but Weka does not implement other models
- Study the use of different fitness functions, mutations, datasets, etc.
- Other uses of GA for SE and PM.