GENETIC PROGRAMMING FOR SOFTWARE DEVELOPMENT EFFORT ESTIMATION

Federica Sarro
fsarro@unisa.it

Outline

1. Background and Motivations
   - Software Development Effort Estimation
   - Effort Estimation with Search-Based Approaches
   - How to Assess Estimation Model Accuracy?

2. Empirical Study: Influence of Fitness Function
   - Research Goals
   - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
   - Results

3. Preliminary Empirical Study: Multi-Objective Genetic Programming
   - Research Goals
   - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
   - Results

4. Conclusions

Obtaining accurate estimates is a critical activity
- for planning and monitoring software project development
- for delivering the product on time and within budget

Significant over or under-estimates expose a software project to several risks
- addition of manpower to a late software project makes the project later (Brooks's Law)
- cancellation of activities, such as documentation and testing, impacts on software quality and maintainability
Software Development Effort Estimation

- Obtaining accurate estimates is a challenging activity
  - the estimation is needed early in the software lifecycle, when little information about the project is available

- Several approaches have been proposed to support project managers in estimating software development effort

Data-Driven Approaches

- A Data-Driven approach exploits data from past projects to estimate the effort for a new project
  - data consist of information about some relevant project features (i.e., cost drivers) and the effort actually spent to develop the projects

Effort Estimation with Search-Based Approaches

- The effort estimation problem can be formulated as an optimization problem
  - we have to find among the possible estimation models the most accurate

- The use of Search-Based (SB) approaches has been suggested for effort estimation
  - the fitness function guides the search
    - it should be able to determine whether an estimation model leads to more accurate predictions than another

---

Effort Estimation with Search-Based Approaches

An SB technique builds many possible models - exploiting past projects data - and tries to identify
the best one
i.e., the one providing the most accurate estimates

An SB technique can be used to improve critical steps of other estimation techniques
e.g., features subset selection or critical parameters setting

SB approaches can be exploited to

build effort estimation models

enhance the use of existing effort estimation methods

Influence of Fitness Function...


How to assess estimation model accuracy?

Several evaluation criteria are employed for assessing the accuracy of effort estimation models

The most commonly used are based on

Absolute Residuals = |ActualEffort - EstimatedEffort|

Summary Measures

- MMRE (Mean MRE)
- MdMRE (Median MRE)
- Pred(25) (Prediction at level 25): percentage of the estimates whose MRE < 25
- MEMRE (Mean EMRE)
- MdEMRE (Median EMRE)

- MRE (Magnitude of Relative Error) \[ MRE = \frac{|ActualEffort - EstimatedEffort|}{ActualEffort} \]
- EMRE (Estimated MRE) \[ EMRE = \frac{|ActualEffort - EstimatedEffort|}{EstimatedEffort} \]

Different accuracy measures take into account different aspects of model performance

- MMRE measures poor performance
- MEMRE is more sensitive to under-estimates
- Pred(25) measures how well an estimation model performs
  - ...
- There is no convergence of opinion on what is the best accuracy measure
  - to compare different models and consistently derive the best one

How to assess estimation model accuracy?

- Different accuracy measures do not account different aspects of estimation accuracy.
  - MMRE measure is monotonic.
  - Pred(25) measure does not penalize underestimates.
  - ... 
- There is no consensus of opinion on what is the best accuracy measure.
  - to compare different models and consistently derive the best one.

What accuracy measure can be used as fitness function?

Some previous works exploited MMRE as fitness function.\(^1,2\)
- one of the most widely used criterion
- one of the most questioned
  - e.g., it does not consistently select the best from two competing models.\(^3\)
- Each measure used to evaluate properties of interest can be used as fitness function.\(^4\)
  - the choice of the evaluation criterion can be a managerial issue
Using Genetic Programming (GP) project managers can select their preferred evaluation criterion as fitness function.
- the search for the estimation model is driven by such a criterion.

Outline

- Background and Motivations
  - Software Development Effort Estimation
  - Effort Estimation with Search-Based Approaches
  - How to Assess Estimation Model Accuracy?
- Empirical Study: Influence of Fitness Function
  - Research Goals
  - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Preliminary Empirical Study: Multi-Objective Genetic Programming
  - Research Goals
  - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Conclusions

---

\(^4\) M. Harman, J.A. Clark, Metrics Are Fitness Functions Too. IEEE METRICS 2004
Empirical Study: Research Goals

- RG₁: How the choice of the fitness function impact on the accuracy of the estimation models built with GP?
  - Does GP effectively optimize the criterion employed as fitness function?
  - Are there any differences in using different fitness functions?
- RG₂: Is GP more effective than widely used effort estimation methods?
  - Manual Stepwise Regression (MSWR), Case-Based Reasoning (CBR), Mean and Median of Effort

Empirical Study: GP Setting (1)

- A solution consists of an estimation model described by an equation
  \[ \text{Effort} = c_1 \cdot f_1 \cdot op_1 \cdot f_2 \cdot \ldots \cdot op_{2n} \cdot f_{2n} \cdot c_{2n} \]
  where
  - \( c_i \) represents the coefficient of the \( i \)th project feature
  - \( f_i \) represents the value of the \( i \)th project feature
  - \( op \in \{+,-,\cdot,\div,\ln(f)\} \)
  - \( C \) represents a constant
  - \( \text{Effort} > 0 \)
- encoded as a binary tree of fixed depth
  - leaves: features and coefficients
  - internal nodes: mathematical operators

Empirical Study: GP Setting (2)

- Initial Population
  - 10V random trees, where \( V \) is the number of project features contained in the dataset
- Genetic Operators
  - crossover randomly selects the same point of cut in parent trees and swaps the corresponding subtrees
  - mutation randomly selects a node in a tree and replaces its value with a new one
- Selection
  - Roulette Wheel Selection for parent selection
  - Tournament Selection for survival selection
- Termination Criteria
  - GP is stopped after 1000V generations or if the fitness value of the best solution does not change after 100V generations
- Execution Number
  - we performed 10 runs considering as final prediction model the one that had the fitness value closest to the average value achieved in the 10 runs on training sets

Empirical Study: GP Setting (3)

- The experimented fitness functions

<table>
<thead>
<tr>
<th>Accuracy Measure</th>
<th>Employed Fitness Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMRE</td>
<td>1/MMRE</td>
</tr>
<tr>
<td>Pred(25)</td>
<td>Pred(25)</td>
</tr>
<tr>
<td>MdMMRE</td>
<td>1/MdMMRE</td>
</tr>
<tr>
<td>MEMRE</td>
<td>1/MEMRE</td>
</tr>
<tr>
<td>MdEMRE</td>
<td>1/MdEMRE</td>
</tr>
<tr>
<td>MMRE e Pred(25)</td>
<td>Pred(25)/MMRE</td>
</tr>
<tr>
<td>MdMMRE e Pred(25)</td>
<td>Pred(25)/MdMMRE</td>
</tr>
<tr>
<td>MEMRE e Pred(25)</td>
<td>Pred(25)/MEMRE</td>
</tr>
<tr>
<td>MdEMRE e Pred(25)</td>
<td>Pred(25)/MdEMRE</td>
</tr>
</tbody>
</table>

We experimented with the above accuracy measures as fitness function to analyze the impact on the estimation accuracy of the constructed models.

The observation that different accuracy measures take into account different aspects of predictions accuracy suggested us to investigate also the effectiveness of some combinations of those accuracy measures.
Empirical Study: Dataset Selection

Table 1. A summary of the employed datasets selected from the PROMISE repository

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Observations</th>
<th>Employed Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desharnais</td>
<td>Software projects derived from a Canadian software house</td>
<td>77</td>
<td>7</td>
</tr>
<tr>
<td>Maxwell</td>
<td>Software projects coming from one of the biggest commercial bank in Finland</td>
<td>62</td>
<td>17</td>
</tr>
<tr>
<td>Telecom</td>
<td>Data about enhancement projects for a U.K. telecommunication product</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>China</td>
<td>Projects developed by Chinese software companies</td>
<td>499</td>
<td>5</td>
</tr>
<tr>
<td>Finnish</td>
<td>Data collected by the TIEKE organizations on projects from different Finnish software companies</td>
<td>38</td>
<td>4</td>
</tr>
<tr>
<td>Kemener</td>
<td>Data on large business applications collected by a national computer consulting and services firm, specialized in the design and development of data-processing software</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Miyazaki</td>
<td>Data on projects developed in 20 companies by Fujitsu Large Systems Users Group</td>
<td>48</td>
<td>3</td>
</tr>
</tbody>
</table>

Empirical Study: Evaluation Criteria

- To assess whether the selected criterion was optimized we employed
  - the summary measure used as fitness function
- To assess the overall estimation accuracy we employed
  - MMRE, Pred(25), MdMRE, MEMRE, MdEMRE
- Boxplots of absolute residuals
- Wilcoxon Test ($\alpha=0.05$) to analyze whether there is significant difference between the absolute residuals
  - since the absolute residuals were not normally distributed and the data was naturally paired

Empirical Study: Validation Method

- We applied a 3-fold cross validation randomly partitioning the original datasets into
  - 3 training sets for model building
  - 3 test sets for model evaluation

Empirical Study: Results Influence of the fitness function (1)

Results on Training Sets...
... to assess the models' ability to fit data
Empirical Study: Results
Influence of the fitness function (1)

Our Running Example is the Desharnais dataset
Note that the observations we will make hold also for the other datasets

Results on Training Sets…
… to assess the models’ ability to fit data

Empirical Study: Results
Influence of the fitness function (3)

Figure 3. Performance of using GP with different fitness functions in terms of MMRE, Pred(25), MdMRE, MEMRE, and MdEMRE on Desharnais dataset (TRAINING SETS)

Empirical Study: Results
Influence of the fitness function (4)

Figure 3. Performance of using GP with different fitness functions in terms of MMRE, Pred(25), MdMRE, MEMRE, and MdEMRE on Desharnais dataset (TRAINING SETS)

Results on Test Sets…
… to assess the models’ predictive capability
Empirical Study: Results
Influence of the fitness function (4)

Our Running Example is again the Desharnais dataset
Note that the observations we will make hold also for the other datasets

Results on Test Sets...
... to assess the models’ predictive capability

Empirical Study: Results
Influence of the fitness function (5)

Figure 5. An excerpt of the trend of summary measures during the evolution process when MMRE is used as fitness function (Desharnais dataset)

Empirical Study: Results
Influence of the fitness function (6)

Figure 6. An excerpt of the trend of summary measures during the evolution process when MdMRE is used as fitness function (Desharnais dataset)
Empirical Study: Results
Influence of the fitness function (7)

Figure 7. Results of the Wilcoxon test comparing different fitness functions (Desharnais dataset)

Each "x" means that "the fitness function indicated on the corresponding row provides significantly less absolute residuals than the fitness function indicated on the corresponding column."


Empirical Study: Results
Influence of the fitness function (8)

Table 2. Influence of the fitness function: a summary

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Best Fitness Functions</th>
<th>Worst Fitness Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desharnais</td>
<td>Pred(25)/MdMRE, Pred(25)/MEMRE</td>
<td>MEMRE, MdEMRE YES NO</td>
</tr>
<tr>
<td>Finnish</td>
<td>Pred(25)/MdMRE, Pred(25)/MEMRE</td>
<td>MEMRE, MdEMRE YES NO</td>
</tr>
<tr>
<td>Kemerer</td>
<td>Pred(25)/MdMRE, Pred(25)/MEMRE</td>
<td>MEMRE, MdEMRE YES NO</td>
</tr>
<tr>
<td>Miyazaki</td>
<td>Pred(25)/MdMRE, Pred(25)/MEMRE</td>
<td>MEMRE, MdEMRE YES NO</td>
</tr>
<tr>
<td>Telecom</td>
<td>Pred(25)/MdMRE, Pred(25)/MEMRE</td>
<td>MEMRE, MdEMRE YES NO</td>
</tr>
<tr>
<td>China</td>
<td>Pred(25)/MdMRE, Pred(25)/MEMRE</td>
<td>MEMRE, MdEMRE YES NO</td>
</tr>
<tr>
<td>Maxwell</td>
<td>Pred(25)/MdMRE, Pred(25)/MEMRE</td>
<td>MEMRE, MdEMRE YES NO</td>
</tr>
</tbody>
</table>

(1) Does using MdMRE as fitness function negatively impact on MEMRE value and viceversa?
(2) Does using MdMRE as fitness function negatively impact on MEMRE value and viceversa?


Empirical Study: Results
GP vs. other estimation methods (1)

Figure 8. Comparison based on Summary Measures
(Desharnais dataset)

- GP achieved the best results in terms of summary measures
- Absolute residuals achieved by GP were significantly less than those achieved by MeanOfEffort, MedianOfEffort, and CBR
- There was not statistical significant difference between GP and MSWR


Empirical Study: Results
GP vs. other estimation methods (2)

Table 3. Wilcoxon Test (p-value)
(Desharnais dataset)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>&lt; MeanOfEffort MeanOfEffort CBR MSWR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desharnais</td>
<td>GP_MdMRE 0 0.002 0.009 0.093</td>
</tr>
<tr>
<td>Finnish</td>
<td>GP_Pred(25) 0 0.001 0.046 0.337</td>
</tr>
<tr>
<td>Miyazaki</td>
<td>GP_Pred(25) 0 0 0.006 0.034</td>
</tr>
<tr>
<td>Maxwell</td>
<td>GP_Pred(25)/MdMRE 0 0.001 0.057 0.691</td>
</tr>
<tr>
<td>Telecom</td>
<td>GP_Pred(25)/MdMRE 0.037 0.01 0.041 0.82</td>
</tr>
<tr>
<td>China</td>
<td>GP_Pred(25)/MdMRE 0 0 0 0.817</td>
</tr>
<tr>
<td>Kemerer</td>
<td>GP_Pred(25)/MdMRE 0.017 0.025 0.295 0.147</td>
</tr>
</tbody>
</table>

How the choice of the fitness function impact on the accuracy of the estimation models built with GP?

GP optimizes the criterion selected as fitness function

- Using MMRE or MEMRE is the worst choice for the overall accuracy
- Using MMRE negatively impacts on MEMRE value and vice versa
- Significantly worse results with respect to the ones achieved using the other fitness functions
- Estimates significantly better than those obtained with CBR
- The fitness functions based on the combination of two criteria often provided better estimates than fitness functions based on a single criterion
- Pred(25)/MMRE, Pred(25)/MdMRE, Pred(25)/MdEMRE can be more promising as fitness function

Empirical Study: Research Goals

- RG1: Is Multi-Objective Genetic Programming effective to address the effort estimation problem?
- RG2: Do the objectives employed in the definition of the fitness function impact on estimation accuracy?
- RG3: Is the increasing of complexity determined by the use of MOGP paid back by an improvement of performance?

Outline

- Background and Motivations
  - Software Development Effort Estimation
  - Effort Estimation with Search-Based Approaches
  - How to Assess Estimation Model Accuracy?
- Empirical Study: Influence of Fitness Function
  - Research Goals
  - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Preliminary Empirical Study: Multi-Objective Genetic Programming
  - Research Goals
  - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Conclusions

Preliminary Empirical Study: MOGP Setting (1)

- We designed and experimented a Multi-Objective Genetic Programming (i.e., MOGP)
- an adaptation to GP of the Non dominated Sort Genetic Algorithm-II (NSGA-II)
- same GP setting except for
  - an objective vector is considered instead of a single function and the fitness assignment is based on the dominance deep according to NSGA-II
  - selection operators perform according to the non-dominance and crowding distance criteria
  - the final solution is selected from the pareto front by using an "a priori" decision maker which provides a complete order between the Pareto optimal solutions according to the following expression
    \[ \text{Pred}(25)/O_1 + ... + O_n \] where \( O_i \) is the value of a measure belonging to the objective vector and to the set (MMRE, MEMRE, MdMRE, MdEMRE)
Different objective vectors were employed as multi-objective functions.

<table>
<thead>
<tr>
<th>Name</th>
<th>Employed Objective Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOGP1</td>
<td>[1/MMRE, Pred(25), 1/MdMRE, 1/MEMRE, 1/MdEMRE]</td>
</tr>
<tr>
<td>MOGP2</td>
<td>[1/MMRE, Pred(25), 1/MdMRE]</td>
</tr>
<tr>
<td>MOGP3</td>
<td>[Pred(25), 1/MMRE]</td>
</tr>
<tr>
<td>MOGP4</td>
<td>[Pred(25), 1/MdMRE]</td>
</tr>
</tbody>
</table>

Validation Method

3-fold cross-validation

Evaluation criteria

summary measures and statistical significance test

Employed Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Observations</th>
<th>Employed Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desharnais</td>
<td>Software projects derived from a Canadian software house</td>
<td>77</td>
<td>7</td>
</tr>
<tr>
<td>Miyazaki</td>
<td>Data on projects developed in 20 companies by Fujitsu Large Systems Users Group</td>
<td>48</td>
<td>3</td>
</tr>
</tbody>
</table>

Influence of the objective vector (1)
Preliminary Empirical Study: Results
Influence of the objective vector (2)

Table 5. p-values of the Wilcoxon test comparing the considered MOGPs (on the test sets)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MOGP1</th>
<th>MOGP2</th>
<th>MOGP3</th>
<th>MOGP4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desharnais</td>
<td>-</td>
<td><strong>0.019</strong></td>
<td>0.124</td>
<td>0.102</td>
</tr>
<tr>
<td>MOGP2</td>
<td>0.981</td>
<td>-</td>
<td>0.927</td>
<td>0.847</td>
</tr>
<tr>
<td>MOGP3</td>
<td>0.876</td>
<td>0.073</td>
<td>-</td>
<td>0.635</td>
</tr>
<tr>
<td>MOGP4</td>
<td>0.898</td>
<td>0.153</td>
<td>0.365</td>
<td>-</td>
</tr>
<tr>
<td>Miyazaki</td>
<td>-</td>
<td><strong>0.033</strong></td>
<td><strong>0.011</strong></td>
<td>-</td>
</tr>
<tr>
<td>MOGP2</td>
<td>0.676</td>
<td>-</td>
<td>0.179</td>
<td>0.142</td>
</tr>
<tr>
<td>MOGP3</td>
<td>0.967</td>
<td>0.821</td>
<td>-</td>
<td>0.971</td>
</tr>
<tr>
<td>MOGP4</td>
<td>0.989</td>
<td>0.858</td>
<td><strong>0.029</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

Null hypothesis: "the use of \( m_i \) does not provide better absolute residuals than using \( m_j \), where \( m_i \) and \( m_j \) are two experimented multi-objective functions"
Outline

- Background and Motivations
  - Software Development Effort Estimation
  - Effort Estimation with Search-Based Approaches
  - How to Assess Estimation Model Accuracy?
- Empirical Study: Influence of Fitness Function
  - Research Goals
  - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Preliminary Empirical Study: Multi-Objective Genetic Programming
  - Research Goals
  - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Conclusions

Conclusions

- GP represents a flexible method that allows project managers to identify their preferred evaluation criterion
  - the choice of the fitness function influences the performance of the models constructed with GP
    - the use of MMRE or MEMRE is not the best choice
      - using them had the effect to degrade a lot of other criteria
    - other accuracy measures are more promising (e.g., Pred(25)/MdMRE)
      - significantly better results than the ones provided by using GP with other fitness functions
      - estimates significantly better than those obtained with CBR
  - A preliminary empirical analysis revealed that
    - the best results achieved with MOGP and GP were comparable
    - the choice of the objective vector influences the performance of the models constructed with MOGP

References


Questions?

Thanks for your attention

Federica Sarro
fsarro@unisa.it
www.dmi.unisa.it/people/sarro/www/

University of Salerno