Reduced-Parameter COCOMO Models

Feature Subset Selection Study of COCOMO in NASA Domain

Zhihao Chen
Tim Menzies
Dan Port
Barry Boehm

Annual Research Review 2005 of Center for Software Engineering
Research Hypothesis and Methods

• Research Hypothesis

  – *For any specific domain and organizational situation there exists a reduced-parameter COCOMO model with higher PRED and lower variance than the general COCOMO model.*
    • A reduced-parameter COCOMO model is the same as the ordinary COCOMO model except uses fewer model calibration features (e.g. effort multipliers)

• Applied model “learner” methods
  – linear model learners
    • Least Squares Regression (LSR) builds one linear model through the entire training data (typical COCOMO calibration method)
    • Model Tree – M5 [6]. M5 can build one or more models. Internally, M5 builds a decision tree whose leaves are linear models which apply to different zones of the parameter space.
  – Feature subset selection [7]
    • Feature subset selection (FSS) is the process of identifying the most promising features in a given dataset.
  – Wrapper [3]
    • WRAPPER is the best FSS mechanism, if the data set is not too large.

• “Holdout” experiments, Wins vs. Losses methods and T-test are used to compare results
  – More on this in later slides

• PRED(N)=P is used to estimate the accuracy of the models.

• Standard Deviation is used to estimate the variability of the models.
Why we use model tree – M5

• Algorithm
  – Internally, M5 builds a decision tree whose leaves are linear models which apply to different zones of the parameter space.
  – M5 computes the standard deviation in the training cases. Then the linear regression models are simplified by dropping attributes if this results in a lower expected error on future data and minimizes its estimated error.

• Benefits
  – In most of the cases, the functional dependency in the COCOMO model is not constant in the whole domain which can be approximated using smaller sub-domains. With M5 model tree, these sub-domains then can be searched for and characterized with some "local" models. These "local" models might be better than a general LSR model.
  – The COCOMO model is simplified by eliminating parameters to minimize its estimated error.
  – A heuristic evaluation function used to select the best partition is computationally efficient. M5 can learn and perform efficiently.
Feature subset selection in Cost Estimation

- What is feature subset selection (FSS)
  - Feature subset selection is the process of identifying the most promising features in a given dataset.

- Why it MAY be useful to effort estimation
  - Feature subset selection helps us identify the important attributes and remove redundant ones.
  - The removal of such features results in a restricted subset of features with equal or better estimation performance than the full feature set.
  - Another benefit is we can use less features to describe the project, and avoid “guessing” uncertain parameters. This enhances the understanding of the dataset or domain under consideration and reduces variability.
  - Additionally, reducing the number of attributes helps us focus on a subset of relevant features, providing the ability to get a better insight in the nature and characteristics of the project.
    - Caution should be practiced here! If an attribute is known to be relevant with little uncertainty then dropping it for the sake of improved estimation on a particular data set may be RISKY.
Example: COCOMO II “Develop for Reuse” Parameter

- **Expert Delphi**
  - Development for multi-domain reuse is 73% more expensive (SD=0.05)

- **COCOMO II Data**
  - Weak dispersion: mostly nominal ratings
  - Noisy data: some don’t-knows reported as nominal
  - Regression Analysis: Development for multi-domain reuse is 17% less expensive (SD=0.28)

- **Bayesian Expert/Data Combination**
  - Development for multi-domain reuse is 31% more expensive
  - Smaller Expert SD gives it heavier weight in averaging

- **Reduced-Parameter Modeling**
  - Machine learner finds better overall COCOMO II accuracy better without RUSE included
  - Better model for organizations not developing for reuse
  - Will add 0% effort to develop for reuse
Hold-Out Experiments

- Linearization

\[
    \text{months} = a \times (K S L O C^b) \times \left( \prod EM_j \right) \quad \Rightarrow \quad \ln(\text{effort}) = b \times \ln(\text{Size}) + \ln(EM_1) + \ln(EM_2) + \ldots
\]

- Evaluation
  - Apply Wrapper using least squares regression and M5 as target learners (Classifier) with 10-fold cross-validation.

- Validation
  - Remove each attribute from lowest rate to highest rate based on the evaluation results
  - The remaining data was randomly sampled 30 times to generate a \(2/3^{rd}\) training set and a \(1/3^{rd}\) test set
  - Least squares regression and M5 model tree were then applied to the training set to learn the linear model of COCOMO 81.
  - This model was then applied to the test set.
  - The mean and standard deviation at PRED(30,40,50) of the model's performance over the 30 subsets was then computed. We call these 30 times experiments as one experiment set.
  - Recall that PRED(30)=50% means that half the estimates are within 30% of the actual.

- Selection
  - We select the feature subsets of COCOMO, which are higher prediction and less standard deviation.
Features Subset Selection Research in Software Cost Estimation Model

The Methods for Identification and Selection

- Four result sets for each experiment set
  - Use \textit{LSR} as target learner to evaluate the cost drivers
    - \textbf{Result One}: Use \textit{least squares regression} approach as linear model validating the feature subsets, and get the result data for \textit{PRED}(30,40,50);
    - \textbf{Result Two}: Use \textit{M5 model tree} approach as linear model validating the feature subsets, and get the result data for \textit{PRED}(30,40,50);
    - We use two approaches (result one and two) to identify and select the \textit{better} feature subsets.
  - Use \textit{M5 model tree} as target learner to evaluate the cost drivers
    - \textbf{Result Three}: Use \textit{least squares regression} approach as linear model validating the feature subsets, and get the result data for \textit{PRED}(30,40,50);
    - \textbf{Result Four}: Use \textit{M5 model tree} approach as linear model validating the feature subsets, and get the result data for \textit{PRED}(30,40,50);
    - We use two approaches (result three and four) to identify and select the \textit{better} feature subsets.

- We repeat such identifications and selections from six experiment sets to find the \textbf{STABLE} better/best feature subsets

- Total \textbf{32400} hold-out experiments for each data set:
  - 30 (hold-out) x 2 (approaches) x 16 (subsets) x 3 (PREDs) x 2 (target learners) x 6 (repeats).
## Results of Selected Feature Subsets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>nasa_ln</th>
<th>Evaluator</th>
<th>Wrapper Subset Evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>60</td>
<td>Evaluation mod 10-fold cross-validation</td>
<td></td>
</tr>
<tr>
<td>Classifier</td>
<td>Linear Squares Regression</td>
<td>Search</td>
<td>BestFirst</td>
</tr>
<tr>
<td>Cost Drivers</td>
<td>SCED RELY LEXP TOOL CPLX MDF PCAP VIRT AEXP DATA VEXP STOR TURN TIME ACAP LOC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation Result</td>
<td>0 0 0 0 0 1 1 1 2 4 7 8 10 10 10 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature Subset</td>
<td>FS01 FS02 FS03 FS04 FS05 FS06 FS07 FS08 FS09 FS10 FS11 FS12 FS13 FS14 FS15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Validating Approach**

<table>
<thead>
<tr>
<th>Selected Feature Subset of LSR as Target Learner</th>
<th>Linear Squares Regression and Model Tree - M5 Prune</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS07</td>
<td>VIRT AEXP DATA VEXP STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>PS09</td>
<td>DATA VEXP STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>FS10</td>
<td>VEXP STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>FS11</td>
<td>STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>FS12</td>
<td>TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>FS13</td>
<td>TIME ACAP LOC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selected Feature Subset of M5 as Target Learner</th>
<th>Linear Squares Regression and Model Tree - M5 Prune</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS06</td>
<td>AEXP DATA CPLX VIRT VEXP STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>PS08</td>
<td>VEXP STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>PS09</td>
<td>VIRT VEXP STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>FS10</td>
<td>VEXP STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>FS11</td>
<td>STOR TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>FS12</td>
<td>TURN TIME ACAP LOC</td>
</tr>
<tr>
<td>Fs13</td>
<td>TIME ACAP LOC</td>
</tr>
</tbody>
</table>

- Methods continue to reduce parameter set until no improvement clearly seen
- Green bars indicate best parameter subsets for each approach
Results of Selected Feature Subsets

Results for LSR as Target Learner

Results for M5 as Target Learner
Some Results

- We apply the methods with LSR as target learner and validating approach on 5 data sets:
  - NASA 60 projects
  - COCOMO81 63 projects (the blue line)
  - NASA Project02 22 projects (the green line)
  - NASA Project03 12 projects
  - NASA Project04 14 projects

- Some observations
  - Highest PRED selected FS not the same for all data sets
  - Highest PRED does not always correspond with lowest SD
  - All PREDs increase then eventually decrease and SD’s always decrease then increase (so cutting all attributes is not good)
Conclusions

• The experiments clearly support the research hypothesis:
  – For any specific domain and organizational situation there exists a reduced-parameter COCOMO model with higher PRED and lower variance than the general COCOMO model.

• But at this point we do not have any means to determine in advance what this reduced model is!

• The significance of cost drivers are different for different projects, different domains and different organizations. While in the same domain and the same organization, the significance of some cost drivers may be similar enough to make reasonable predictions.

• In the same domain and the same organization, some attributes may be correlated or irrelevant. Correlated or irrelevant attributes often behave like “noise” in the data, confusing the cost model and degrading its performance. The removal of such attributes results in a restricted subset of features with equal or better estimation performance than the full feature set.

• Another benefit is we can use less features to describe the project, get a better insight in the nature and characteristics of the project, and to make a faster estimation.

• However, a dropped parameter means that changes in this parameter have no impact on estimated costs. This is OK if the parameter is irrelevant to the organization, but not otherwise.
Threats to External Validity

- To avoid the threat of order effects
  - Randomly select the training set and test set;
  - Repeat the hold-out experiment 30 times;
  - Repeat the above experiment set six times.

- Randomization removes bias

- To avoid the chance of over-fitting
  - Conduct holdout studies – 2/3 as training set and 1/3 as test set.

- To find the STABLE selected feature subsets
  - Repeat the whole experiments six times;
  - The selected feature subsets get higher predictions and lower standard deviations at PRED(30,40,50);
  - Validate the results with two target learners;
  - Validate the results with two approaches.
Future Work

• The current work is in NASA domain with COCOMO 81 model.

• We are going to navigate the some domains / organizations with COCOMO II model.
References


• Estimate Model Accuracy

• What is variability in cost modeling?

• Paired t-test in our model
Estimate Model Accuracy

• What is the accuracy of the estimation?
  – It defines the closeness of estimation to the actual cost
  – More robust than sum of squared residuals or mean squared error

\[
RE_i = \frac{\text{estimate}_i - \text{actual}_i}{\text{actual}_i} \quad MRE_i = \text{abs}(RE_i)
\]

\[
PRED(N) = \frac{100}{T} \sum_{i}^{T} \begin{cases} 
1 & \text{if } MRE_i \leq \frac{N}{100} \\
0 & \text{otherwise}
\end{cases}
\]

• PRED(N)=P
  – higher P, lower N implies greater accuracy

• Some basic questions about use of PRED. For any given estimate not used in the calibration:
  – What is the likelihood that this estimate is in the P%?
  – How “far out” from the N% might this estimate be if not within the P%?
  – ** How confident are we that the estimate it is represented with this PRED?

Example: PRED(30) = 50 Means that half the estimations are within 30% of the actual.
What is variability in cost modeling?

- Accuracy is not enough to answer the basic questions. Need to consider variability.
  - PRED(N) is calculated for different “holdout” samplings of the calibration data.
    - Holdouts use randomized sub-samples of the data to calibrate and the un-sampled data to calculate PRED.
    - Different samplings produce different PRED(N) values. $\mu$, is the mean of PRED(N) of these.
  - Variability shows how much spread is in the estimation. Standard deviation $\sigma$ is used to measure variability.
  - In figure 2 smaller $\sigma$ provides more “confidence” in using PRED(N) = $\mu$ than figure 1 since all the values are closer to $\mu$.
  - Small $\sigma$ indicates small variability in estimations.
  - Large $\sigma$ indicates large variability in estimation.

- Mean and Standard Deviation:

\[
\mu(x) = \frac{1}{N} \sum_{i=1}^{N} x_i \quad V = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \quad \sigma = \sqrt{V}
\]
Paired t-test in our model

- T-test is a statistical test which allows one to say something about differences between means at a certain confidence level.

- Null hypothesis of the T-test: no difference exists between the means.

- We use it in our model to determine the selected feature subsets.

- Possible results among the full feature set and the feature subsets:
  - I am 99% sure that null hypothesis is rejected between the means of these two subsets.
    - There is probably a true difference between the means:
      - This mean in this subset is larger than that in that subset; this subset wins;
      - This mean in this subset is smaller than that in that subset; this subset losses;
    - I cannot reject the null hypothesis the means are likely the same.

- The number of degrees of freedom (df) = N – 1, where N = 30 is the number in hold-out experiments. Confident level is 0.05.