Software Effort Estimation as a Multi-objective Learning Problem

Leandro Minku (www.cs.bham.ac.uk/~minkull)

CERCIA, School of Computer Science, The University of Birmingham

January 31, 2013
ML models for Software Effort Estimation (SEE).
  - Decision support tools.
Introduction – ML and ensembles

ML models for Software Effort Estimation (SEE).
- Decision support tools.

**Ensembles** of learning machines:
- Recently attracted attention of SEE community.
- Tailoring is necessary (**base** learner choice or ensemble method).


Base learners in ensembles should be diverse.

Different performance measures for evaluating SEE can behave differently.

- MMRE, PRED, LSD, MAE, etc.
Base learners in ensembles should be diverse.

Different performance measures for evaluating SEE can behave differently.

- MMRE, PRED, LSD, MAE, etc.

**Question**

Can we use that for improving SEEs?

L. Minku and X. Yao. Software effort estimation as a multi-objective learning problem. TOSEM (accepted).
Base learners in ensembles should be diverse.

Different performance measures for evaluating SEE can behave differently.

- MMRE, PRED, LSD, MAE, etc.

**Question**

Can we use that for improving SEEs?

L. Minku and X. Yao. Software effort estimation as a multi-objective learning problem. TOSEM (accepted).

1. How differently do these measures behave in SEE?
Base learners in ensembles should be diverse.

Different performance measures for evaluating SEE can behave differently.

- MMRE, PRED, LSD, MAE, etc.

**Question**

1. How differently do these measures behave in SEE?
2. Can we use them to create good ensembles for SEE?

L. Minku and X. Yao. Software effort estimation as a multi-objective learning problem. TOSEM (accepted).
Base learners in ensembles should be **diverse**.

Different performance measures for evaluating SEE can behave differently.

- MMRE, PRED, LSD, MAE, etc.

**Question**

1. How differently do these measures behave in SEE?
2. Can we use them to create good ensembles for SEE?
3. Can we emphasize a particular measure if we wish to?
Learn models for SEE.

Each performance measure is an objective to be optimised.


Multi-Objective Evolutionary Algorithm:
  - Can be used for answering our research questions.
MOEAs are population-based optimisation algorithms.

Multiple-objectives, possibly conflicting – dominance:

\[ f_i(x^{(1)}) \leq f_i(x^{(2)}) \quad \forall i \land \exists i \mid f_i(x^{(1)}) < f_i(x^{(2)}) \]

“Pareto solutions” – nondominated solutions in the last generation, generally good at all objectives.

Solutions should be diverse, spread well over the objective space.
Performance measures for creating models:

- Mean Magnitude of the Relative Error:

  \[ MMRE = \frac{1}{T} \sum_{i=1}^{T} MRE_i, \]

  where \( MRE_i = |\hat{y}_i - y_i|/y_i; \) \( \hat{y}_i \) is the predicted effort; and \( y_i \) is the actual effort.

- Percentage of estimations within 25\% of the actual values:

  \[ PRED(25) = \frac{1}{T} \sum_{i=1}^{T} \begin{cases} 1, & \text{if } MRE_i \leq \frac{25}{100} \\ 0, & \text{otherwise} \end{cases}. \]

- Logarithmic Standard Deviation:

  \[ LSD = \sqrt{\frac{\sum_{i=1}^{T} \left( e_i + \frac{s^2}{2} \right)^2}{T - 1}}, \]

  where \( s^2 \) is an estimator of the variance of the residual \( e_i \) and \( e_i = \ln y_i - \ln \hat{y}_i. \)
MOEA: Harmonic Distance MOEA.

Objectives/performance: calculated on training set.

SEE Models: Multi-Layer Perceptrons (MLPs).

Representation: vector of real values (weights and thresholds).

Crossover: \( w^c = w^{p1} + N(0, \sigma^2)(w^{p2} - w^{p3}) \)

Self-tuning crossover: \( \sigma^2 = 2 - \left( \frac{1}{1 + e^{(\text{anneal time} - \text{generation})}} \right) \)

Mutation: \( w_i = w_i + N(0, 0.1) \)

Optional: training with Backpropagation.


Two different ways to use solutions:

- Ensemble of “best fit” Pareto solutions:
  - Ensemble SEE = average SEE of base models.
  - Good trade-off among measures.

- Use one best fit Pareto solution.
Experiments

- Data sets: cocomo81, nasa93, nasa, cocomo2, desharnais, 7 ISBSG organization type subsets.
  - ISBSG subsets’ productivity rate is statistically different.
  - Attributes: cocomo attributes + loc for PROMISE data, functional size, development type and language type for ISBSG.
Experiments

- Data sets: cocomo81, nasa93, nasa, cocomo2, desharnais, 7 ISBSG organization type subsets.
  - ISBSG subsets’ productivity rate is statistically different.
  - Attributes: cocomo attributes + loc for PROMISE data, functional size, development type and language type for ISBSG.

- 30 runs for each data set, test in a holdout set with 10 projects.
Experiments

- Data sets: cocomo81, nasa93, nasa, cocomo2, desharnais, 7 ISBSG organization type subsets.
  - ISBSG subsets’ productivity rate is statistically different.
  - Attributes: cocomo attributes + loc for PROMISE data, functional size, development type and language type for ISBSG.

- 30 runs for each data set, test in a holdout set with 10 projects.

- Performance measures for evaluation on test set: MMRE, PRED(25), LSD, MdMRE, MAE, MdAE.
Experiments

- Data sets: cocomo81, nasa93, nasa, cocomo2, desharnais, 7 ISBSG organization type subsets.
  - ISBSG subsets’ productivity rate is statistically different.
  - Attributes: cocomo attributes + loc for PROMISE data, functional size, development type and language type for ISBSG.

- 30 runs for each data set, test in a holdout set with 10 projects.

- Performance measures for evaluation on test set: MMRE, PRED(25), LSD, MdMRE, MAE, MdAE.

- Effect size: $\frac{|M_a - M_p|}{SD_p}$

Experiments

- Data sets: cocomo81, nasa93, nasa, cocomo2, desharnais, 7 ISBSG organization type subsets.
  - ISBSG subsets’ productivity rate is statistically different.
  - Attributes: cocomo attributes + loc for PROMISE data, functional size, development type and language type for ISBSG.

- 30 runs for each data set, test in a holdout set with 10 projects.

- Performance measures for evaluation on test set: MMRE, PRED(25), LSD, MdMRE, MAE, MdAE.

- Effect size: \( \frac{|M_a - M_p|}{SD_p} \)


- Comparing approaches:
  - MLP, RBF;
  - REPTree, Bagging+MLP, Bagging+REPTree, log + EBA;
  - Bagging+RBF, Rand+MLP, NCL+MLP.
Research Question 1

How differently do the performance measures behave in SEE? (Are they different enough for using them as a source of diversity in ensembles?)

- MMRE, PRED(25), LSD.
The Relationship Among Different Performance Measures

Example of Pareto solutions for Cocomo 81.

- More different behaviour than one may have first thought.
- Choosing may still not be easy, so we propose our ensemble approach, which automatically provides a good trade-off among measures.

Leandro Minku (www.cs.bham.ac.uk/~minkull)
Research Question 2

Can we use different performance measures to create good ensembles for SEE?

- Can it improve an MLP on the performance measures used as objectives?
- Can it improve on other approaches (mixed evaluation of MOEA and MLP)?
- And what about other performance measures?
### Results for large (≥ 60) data sets:

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Pareto Ensemble</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LSD</td>
<td>MMRE</td>
<td>PRED(25)</td>
<td></td>
</tr>
<tr>
<td>Wins</td>
<td>6/8</td>
<td>5/8</td>
<td>7/8</td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>0.0000</td>
<td>0.0012</td>
<td>0.0003</td>
<td></td>
</tr>
</tbody>
</table>

### Results for small (< 35) data sets:

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Pareto Ensemble</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LSD</td>
<td>MMRE</td>
<td>PRED(25)</td>
<td></td>
</tr>
<tr>
<td>Wins</td>
<td>3/5</td>
<td>2/5</td>
<td>3/5</td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>0.1170</td>
<td>0.7166</td>
<td>0.0004</td>
<td></td>
</tr>
</tbody>
</table>
Research Question 2

Question 2

Can we use different performance measures to create good ensembles for SEE?

- Can it improve an MLP on the performance measures used as objectives?
  Yes, similar or better performance was obtained across data sets on all objectives. It is worth considering objectives explicitly.
Question 2

Can we use different performance measures to create good ensembles for SEE?

- Can it improve an MLP on the performance measures used as objectives?
  Yes, similar or better performance was obtained across data sets on all objectives. It is worth considering objectives explicitly.

- Can it improve on other approaches (mixed evaluation of MOEA and MLP)?

- And what about other performance measures?
Comparison Against Other Approaches

Performance measures: LSD, MMRE, PRED(25), MdMRE, MAE, MdAE.

Friedman test: models are different across data sets.
- Top half ranked approaches (except for LSD):
  - Pareto ensemble, bagging + MLP, log + EBA, RTs.
  - Pareto ensemble and log + EBA have median ranking standard deviation.
  - Models based on MLPs do not perform well on LSD – negative estimations.
  - MOEAs could be used to evolve other types of model.
### Comparison Against Other Approaches

#### Best ranked approach for each data set:

<table>
<thead>
<tr>
<th>Approach</th>
<th>LSD</th>
<th>MMRE</th>
<th>PRED(25)</th>
<th>MdMRE</th>
<th>MAE</th>
<th>MdAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cocomo81</td>
<td>RT</td>
<td>Bag+MLP</td>
<td>Bag+MLP</td>
<td>Bag+ MLP</td>
<td>Bag + MLP</td>
<td>Bag + MLP</td>
</tr>
<tr>
<td>Sdr</td>
<td>RT</td>
<td>Bag+MLP</td>
<td>Bag+MLP</td>
<td>Bag + MLP</td>
<td>Bag + MLP</td>
<td>Bag + MLP</td>
</tr>
<tr>
<td>Nasa</td>
<td>Bag+RT</td>
<td>RT</td>
<td>Bag+MLP</td>
<td>Bag+MLP</td>
<td>Bag + MLP</td>
<td>Bag + MLP</td>
</tr>
<tr>
<td>Desharnais</td>
<td>Bag+RT</td>
<td>Bag+MLP</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
</tr>
<tr>
<td>Nasa93</td>
<td>Bag+RT</td>
<td>RT</td>
<td>Bag+MLP</td>
<td>Bag+MLP</td>
<td>Bag + MLP</td>
<td>Bag + MLP</td>
</tr>
<tr>
<td>Org1</td>
<td>Bag+RBF</td>
<td>Bag+MLP</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
</tr>
<tr>
<td>Org2</td>
<td>Bag+RT</td>
<td>Bag+MLP</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
</tr>
<tr>
<td>Org3</td>
<td>Pareto Ens</td>
<td>Bag+MLP</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
</tr>
<tr>
<td>Org4</td>
<td>Bag+RBF</td>
<td>Bag+MLP</td>
<td>Log + EBA</td>
<td>Log + EBA</td>
<td>Log + EBA</td>
<td>Log + EBA</td>
</tr>
<tr>
<td>Org5</td>
<td>Bag+RT</td>
<td>Bag+RBF</td>
<td>Bag+RBF</td>
<td>Bag + MLP</td>
<td>Bag + MLP</td>
<td>Bag + MLP</td>
</tr>
<tr>
<td>Org6</td>
<td>Bag+RBF</td>
<td>Bag+MLP</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
</tr>
<tr>
<td>Org7</td>
<td>Bag+RT</td>
<td>Bag+MLP</td>
<td>Log + EBA</td>
<td>Log + EBA</td>
<td>Log + EBA</td>
<td>Log + EBA</td>
</tr>
<tr>
<td>OrgAll</td>
<td>Bag+RBF</td>
<td>Bag+MLP</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
<td>Pareto Ens</td>
</tr>
</tbody>
</table>

Pareto ensemble was ranked first more often for the ISBSG data sets.

Possible reason: MOEA performs global optimisation. More heterogeneous data sets may present several peaks.
Comparison Against Other Approaches

Number of times ranked best:

<table>
<thead>
<tr>
<th>Approach</th>
<th>LSD</th>
<th>MMRE</th>
<th>PRED(25)</th>
<th>MdMRE</th>
<th>MAE</th>
<th>MdAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto Ens</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>RT</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Bag+RT</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Bag+MLP</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Log + EBA</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bag+RBF</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Rand+MLP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RBF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

Pareto ensemble is more often ranked first than other approaches, except for LSD.
Comparison Against Other Approaches

Number of times ranked worst:

<table>
<thead>
<tr>
<th>Approach</th>
<th>LSD</th>
<th>MMRE</th>
<th>PRED(25)</th>
<th>MdMRE</th>
<th>MAE</th>
<th>MdAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag + MLP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MLP</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Bag + RT</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pareto Ens</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Rand + MLP</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bag + RBF</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>RBF</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>NCL + MLP</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

Pareto ensemble is never ranked worst more than twice.
Comparison Against Other Approaches

Effect size against Pareto ensemble in terms of MAE:

<table>
<thead>
<tr>
<th>Approach</th>
<th># Small</th>
<th># Medium</th>
<th># Large</th>
<th># Medium+Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag+MLP</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Bag+RBF</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Bag+RT</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Log + EBA</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>MLP</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>NCL + MLP</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Rand + MLP</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>RBF</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>RT</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Choosing between Pareto ensemble and other approach results in many medium or large effect sizes, representing a considerable practical impact.
<table>
<thead>
<tr>
<th>Question 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can we use different performance measures to create good ensembles for SEE?</td>
</tr>
<tr>
<td>- Can it improve an MLP on the performance measures used as objectives?</td>
</tr>
<tr>
<td>- Can it improve on other approaches (mixed evaluation of MOEA and MLP)? Yes. Pareto ensemble was frequently ranked first and rarely ranked worst, having median stability and being helpful especially for more heterogeneous data sets.</td>
</tr>
<tr>
<td>- And what about other performance measures? The statistics show that the Pareto ensemble is competitive considering all measures but LSD.</td>
</tr>
</tbody>
</table>
Research Question 3

Question 3

Can we emphasize a particular measure if we wish to?

Yes. Using the best fit Pareto solution in terms of a performance measure provides similar or better performance in terms of this measure, but similar or worse in terms of the other measures.

Work is robust to new findings.
Conclusions

- We view the problem of creating SEE models as a multi-objective learning problem.
- We showed to what extent different performance measures behave differently.
- Using a Pareto ensemble of MLPs improved results in terms of all objectives against traditional MLPs.
- The Pareto ensemble of MLPs was competitive against other approaches.
- It is also possible to emphasize a certain performance measure if desired.
Future Work

- Pareto ensemble did better for more heterogeneous data sets.
  - Recent results showing that cross-company data can improve within-company SEEs.
  - When can we learn from other companies? When to change our models?


- MOEAs could also be used to create other types of base model than MLPs – can we improve by creating local models?
- A further study of the choice of Pareto solutions to include in the ensemble showed that there is still room for improvement.
- Different MOEAs could be investigated.

L. Minku and X. Yao. Software effort estimation as a multi-objective learning problem. TOSEM (accepted)
http://www.cs.bham.ac.uk/~minkull/publications