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

Introduction – ML and ensembles

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).

E. Kocaguneli, T. Menzies and J. Keung. On the value of ensemble effort estimation. TSE in press.

L. Minku and X. Yao. Ensembles and locality: insight on improving software effort estimation. IST in press.

Introduction – diversity and performance measures

Base learners in ensembles should be diverse.

Different performance measures for evaluating SEE can behave differently.

- MMRE, PRED, LSD, MAE, etc.

Introduction – diversity and performance measures

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).

Introduction – diversity and performance measures

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?

Introduction – diversity and performance measures

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?
2. Can we use them to create good ensembles for SEE?

Introduction – diversity and performance measures

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?
2. Can we use them to create good ensembles for SEE?
3. Can we emphasize a particular measure if we wish to?

SEE as a Multi-Objective Learning Problem

- Learn models for SEE.
- Each performance measure is an objective to be optimised.

  M. Harman and J. Clark. Metrics are fitness functions too. METRICS 2004.
- Multi-Objective Evolutionary Algorithm:
  - Can be used for answering our research questions.

Multi-Objective Evolutionary Algorithms (MOEAs)

- MOEAs are population-based optimisation algorithms.
- Multiple-objectives, possibly conflicting – dominance:

$$f_i(x^{(1)}) \leq f_i(x^{(2)}) \; \forall i \wedge \exists i \mid f_i(x^{(1)}) < f_i(x^{(2)})$$

- *"Pareto solutions"* – nondominated solutions in the last generation, generally good at all objectives.
- Solutions should be diverse, spread well over the objective space.

Using MOEAs for Creating SEE Models

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 $e_i = \ln y_i - \ln \hat{y}_i$.

Using MOEAs for Creating SEE Models

- 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: $w^c = w^{p_1} + N(0, \sigma^2)(w^{p_2} - w^{p_3})$
- 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.

Z. Wang, K. Tang and X. Yao. Multi-objective approaches to optimal testing resource allocation in modular software systems. TR, 2010.

A. Chandra and X. Yao. Ensemble learning using multi-objective evolutionary algorithms. JMMA, 2006.

Using MOEAs for Creating SEE Models

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

M. Shepperd and S. MacDonell. Evaluating prediction systems in software project estimation. IST 2012.

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

  M. Shepperd and S. MacDonell. Evaluating prediction systems in software project estimation. IST 2012.
- Comparing approaches:
  - MLP, RBF;
  - REPTree, Bagging+MLP, Bagging+REPTree, log + EBA;
  - Bagging+RBF, Rand+MLP, NCL+MLP.

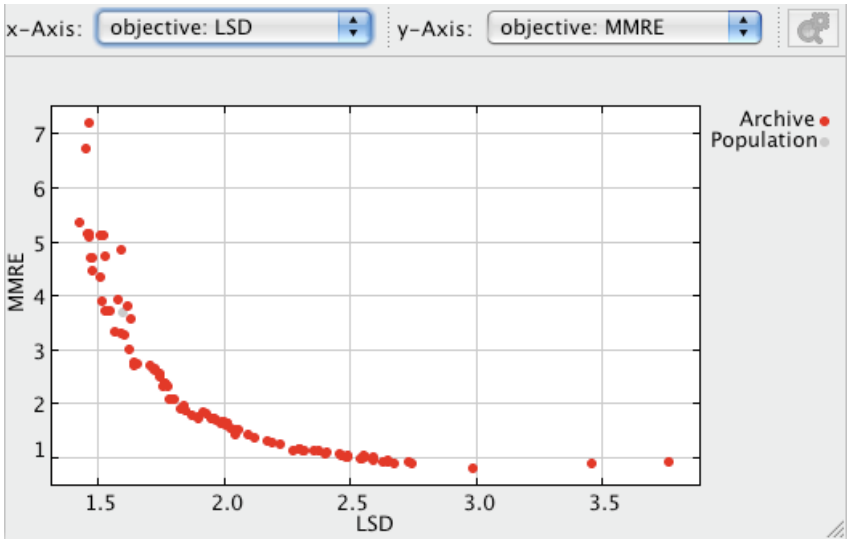
Research Question 1

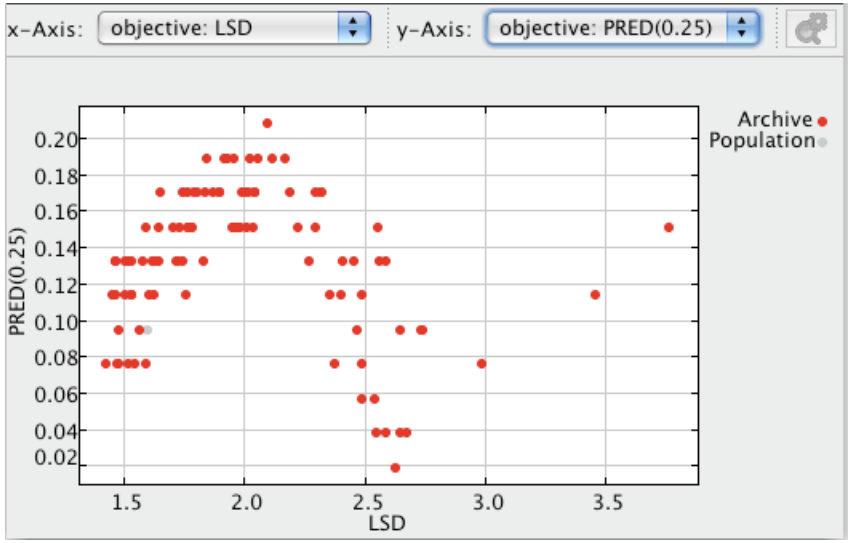
Question 1

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.

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?
- Can it improve on other approaches (mixed evaluation of MOEA and MLP)?
- And what about other performance measures?

Pareto Ensemble Vs Backpropagation MLP

Results for large (> 60) data sets:

Data Set	Pareto Ensemble		
	LSD	MMRE	PRED(25)
Wins	6/8	5/8	7/8
P-value	0.0000	0.0012	0.0003

Results for small (< 35) data sets:

Data Set	Pareto Ensemble		
	LSD	MMRE	PRED(25)
Wins	3/5	2/5	3/5
P-value	0.1170	0.7166	0.0004

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.

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.

- 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:

Approach	LSD	MMRE	PRED(25)	MdMRE	MAE	MdAE
Cocomo81	RT	Bag+MLP	Bag+MLP	Bag+ MLP	Bag + MLP	Bag + MLP
Sdr	RT	RT	Bag+RT	RT	RT	RBF
Nasa	Bag+RT	RT	Bag+MLP	Bag + MLP	Bag +RT	Bag + RT
Desharnais	Bag+RT	Bag+MLP	Pareto Ens	Pareto Ens	Pareto Ens	Pareto Ens
Nasa93	RT	RT	RT	RT	RT	RT
Org1	Bag+RBF	Pareto Ens	Pareto Ens	Pareto Ens	Pareto Ens	Pareto Ens
Org2	Bag+RT	Pareto Ens	Pareto Ens	Pareto Ens	Pareto Ens	Pareto Ens
Org3	Pareto Ens	Pareto Ens	Log + EBA	Log + EBA	Log + EBA	Log + EBA
Org4	Bag+RBF	Pareto Ens	RT	RT	Pareto Ens	Pareto Ens
Org5	Bag+RT	Log + EBA	Bag+RBF	Rand + MLP	Bag + RT	RT
Org6	Bag+RBF	Pareto Ens	Pareto Ens	Pareto Ens	Bag + RBF	Pareto Ens
Org7	Bag+RT	Log + EBA	Log + EBA	Log + EBA	Bag + RBF	Pareto Ens
OrgAll	RT	Pareto Ens	Pareto Ens	Pareto Ens	Pareto Ens	Pareto Ens

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:

Approach	LSD	MMRE	PRED(25)	MdMRE	MAE	MdAE
Pareto Ens	1	6	5	5	5	7
RT	4	3	2	3	2	2
Bag+RT	5	0	1	0	2	1
Bag+MLP	0	2	2	2	1	1
Log + EBA	0	2	2	2	1	1
Bag+RBF	3	0	1	0	2	0
Rand+MLP	0	0	0	1	0	0
RBF	0	0	0	0	0	1
Total	13	13	13	13	13	13

Pareto ensemble is more often ranked first than other approaches, except for LSD.

Comparison Against Other Approaches

Number of times ranked worst:

Approach	LSD	MMRE	PRED(25)	MdMRE	MAE	MdAE
Bag + MLP	0	0	0	0	1	0
MLP	1	0	1	0	0	0
RT	0	0	0	1	1	0
Bag + RT	0	1	1	0	0	1
Pareto Ens	1	2	0	1	1	1
Rand + MLP	2	1	1	2	1	1
Bag + RBF	0	3	3	2	0	2
RBF	1	2	4	3	4	3
NCL + MLP	8	4	3	4	5	5
Total	13	13	13	13	13	13

Pareto ensemble is never ranked worst more than twice.

Comparison Against Other Approaches

Effect size against Pareto ensemble in terms of MAE:

Approach	# Small	# Medium	# Large	# Medium+Large
Bag+MLP	7	4	2	6
Bag+RBF	6	3	4	7
Bag+RT	7	3	3	6
Log + EBA	7	2	4	6
MLP	7	3	3	6
NCL + MLP	5	3	5	8
Rand + MLP	7	3	3	6
RBF	6	2	5	7
RT	4	6	3	9

Choosing between Pareto ensemble and other approach results in many medium or large effect sizes, representing a considerable practical impact.

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?
- 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.
- And what about other performance measures? The statistics show that the Pareto ensemble is competitive considering all measures but LSD.

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?

    L. Minku and X. Yao. Can Cross-company Data Improve Performance in Software Effort Estimation?, PROMISE 2012.
- 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`