Empirical Evaluation of Analogy-X for Software Cost Estimation

Jacky Keung
Empirical Software Engineering Research Group, NICTA Ltd.
CSE, University of New South Wales
Sydney, Australia
Jacky.Keung@nicta.com.au

ABSTRACT
This paper reports on the empirical evaluation of a novel approach called Analogy-X, which is an extension to the classical analogy-based software cost estimation. The Analogy-X approach is a set of procedures that utilize the principles of the Mantel randomization test to provide inferential statistics to Analogy. Our previous studies have clearly demonstrated the novelty and effectiveness of this technique. This paper provides further empirical evaluation of Analogy-X using different kinds of datasets. Our results show that the prediction accuracy of Analogy-X is similar to the one of ANGEL. Analogy-X has the additional advantage of allowing the use of Mantel statistics to select project features and detect abnormal data points, which provides a sound statistical basis for analogy-based systems.

Categories and Subject Descriptors
D.2.9 [**Software Engineering**]: Management – *Cost Estimation*

General Terms
Management, Measurement, Experimentation

Keywords
Software Measurement, Software Effort Estimation, Analogy

1. THE ANALOGY-X APPROACH
The basic principle underlying analogy-based software project effort estimation is that:

"Projects that are similar with respect to project features will be also similar with respect to project effort."

However it is not possible to formally test this hypothesis with the traditional analogy-based approach as described in Shepperd et al. [1]. We have proposed a new statistical approach Analogy-X (Analogy eXtension) in [2][3], which provides a means of formally testing this hypothesis using Mantel's randomization test for correlating two distance matrices. In our previous studies, the results were promising, showing Analogy-X can be used to assess whether analogy is indeed appropriate for the given dataset or not. An approach similar to stepwise regression analysis has been developed in [2] to support feature selection based on statistics rather than on brute-force as in a traditional analogy-based system [1]. We also developed a leverage metric to support sensitivity analysis [2], ensuring the dataset is error-free and preventing the introduction of spurious correlation.

2. THE EXPERIMENT
In our previous studies [2][3], we have only examined the effect and result using three large to ultra large datasets, Desharnais, CSC and ISBSG datasets [1][2]. In this study, the intention is to apply Analogy-X to three smaller datasets which have been used in [1], Albrecht, Telecom1, and Kemerer. The purpose is to assess that the Analogy-X approach is also applicable to a wider range of datasets. Especially, the size of these datasets is relatively small compared to the large datasets used in [2][3]. In reality, most of the SME software organizations would have only a small amount of project cases for software cost estimation. This study also has a focus on the application of Analogy-X on more homogenized datasets within the same software organization.

The evaluation criterion here is that Analogy-X should be capable of statistically testing the above stated basic principle of analogy in section 1. That is to use Analogy-X to check whether or not the dataset under investigation can be successfully applied with analogy-based prediction systems. The empirical assessment of Analogy-X is based on its conformance with the known expected outcomes under various conditions given by different datasets in the study.

The statistical procedures of Analogy-X used in the following experiments have been described in [2][3].

3. ALBRECHT DATASET
The Albrecht dataset contains 24 completed software project cases carried out in IBM in the 1970's. The project features of this dataset are depicted in the following table (the response variable is ActualEffort):

Independent Variable	Description	Type	Mantel-R	p-value
Input	No. of Inputs	Numerical	0.5781	0.160
Output	No. of Outputs	Numerical	0.7988	0.001
Inquiry	No. of Inquires	Numerical	0.8216	0.001
File	No. of Files	Numerical	0.6613	0.001
FPAdj	Adjustment Factor	Numerical	-0.068	0.741
RawFPCount	Raw Function Points	Numerical	0.9370	0.001
AdjFP	Adjusted Function Points	Numerical	0.9111	0.001

Table 1. Project Features of Albrecht Dataset

ESEM'08, October 9–10, 2008, Kaiserslautern, Germany.

In the first stepwise iteration, Analogy-X successfully identified the variable *RawFPCount* being the most useful feature among all the variables considered. With a Mantel's correlation value R = 0.9370 and p-value = 0.001 (see Table 1) *RawFPCount* is significantly correlated with *ActualEffort*. The Jackknife estimator of R is 0.9362 and its upper confidence limit is 0.9538 (the procedure to derive Jackknife estimator values can be found in [2] and [3]). Additional variables in combination with *RawFPCount* must exceed 0.9538. However, no further combination of *RawFPCount* with another variable improved the Mantel correlation. Table 2 shows the project feature selected using Analogy-X.

Iteration	Feature subset	Mantel R	p-value	Jackknife R	UCL
1	RawFPcount	0.93698	0.001	0.93622	0.95377
	➩ Final Model				

Table 2 Project feature selected using Analogy-X

The entire stepwise operation completed in less than 3 seconds. The Analogy-X sensitivity analysis procedure is also performed to ensure the selected feature is not an artifact of any abnormal projects.

It is interesting to note that ANGEL[1]'s brute-force feature selection identified four features which are different to the ones identified by Analogy-X above, these features are *Input, Inquiry, File and AdjFP* with *MMRE* = 0.647 (Mean Magnitude Relative Error [4]). We use the validation functionality of ANGEL to validate the most useful feature *RawFPCount* identified by Analogy-X, the *MMRE* value of 1.006 is worse than the *MMRE* value of 0.647 by using ANGEL (Note: The smaller the *MMRE* the better the prediction will be [4]).

However the measure of *MMRE* is technically unreliable as explained by Foss et al. [4]. Further investigation using a paired t-test is performed to compare the absolute residual between the predictions produced by ANGEL and Analogy-X. In this case, the alternative hypothesis is that ANGEL's prediction residual (absolute difference) values are significantly less than the prediction residual values of Analogy-X's prediction. Otherwise both predictions are similar.

Table 3 shows the *actual* and *predicted* values of *effort* and the residual of both ANGEL and Analogy-X. Based on the predictions using Analogy-X, the mean of absolute residual difference is 1.5, the paired t-test between ANGEL and Analogy-X is 0.5431 with a p-value 0.7039 for the alternative hypothesis. This suggests that the difference between predictions produced by ANGEL and the predictions produced by Analogy-X is statistically insignificant, and there is sufficient evidence to show that both predictions using different feature subsets are very similar.

Case	Actual Effort	ANGEL Prediction	Absolute Residual	Analogy-X Prediction	Absolute Residual	Absolute Residual Difference
1	102.4	61.2	41.2	105.2	2.8	38.4
2	105.2	61.2	44	102.4	2.8	41.2
3	11.1	8	3.1	18.3	7.2	-4.1
4	21.1	19	2.1	19	2.1	0
5	28.8	18.3	10.5	3.6	25.2	-14.7
6	10	6.1	3.9	8.9	1.1	2.8
7	8	2.9	5.1	4.9	3.1	2
8	4.9	4.1	0.8	8	3.1	-2.3
9	12.9	19	6.1	11.8	1.1	5
10	19	21.1	2.1	11.8	7.2	-5.1
11	10.8	8.9	1.9	3.6	7.2	-5.3
12	2.9	0.5	2.4	6.1	3.2	-0.8
13	7.5	15.8	8.3	15.8	8.3	0
14	12	3.6	8.4	12.9	0.9	7.5
15	4.1	4.9	0.8	0.5	3.6	-2.8
16	15.8	7.5	8.3	7.5	8.3	0
17	18.3	21.1	2.8	11.1	7.2	-4.4
18	8.9	10.8	1.9	10	1.1	0.8
19	38.1	61.2	23.1	61.2	23.1	0
20	61.2	38.1	23.1	38.1	23.1	0
21	3.6	10	6.4	28.8	25.2	-18.8
22	11.8	12.9	1.1	19	7.2	-6.1
23	0.5	2.9	2.4	4.1	3.6	-1.2
24	6.1	10	3.9	4.9	1.2	2.7
Mean	**21.9**	**17.9**	**8.9**	**21.6**	**7.5**	**1.5**

Table 3 Predictions using ANGEL and Analogy-X

4. TELECOM1 DATASET

A similar procedure was applied to the Telecom1 dataset [1]. There are 18 enhanced software products in the Telecom1 dataset, Table 4a shows the features of the Telecom1 dataset.

Independent Variable	Description	Type	Mantel-R	p-value
CHANGES	No. of changes	Numerical	0.5286	0.002
FILES	No. of files	Numerical	0.4676	0.003

Iteration	Feature subset	Mantel R	p-value	Jackknife R	UCL
1	Changes	0.5286	0.002	0.5244	0.66
	➩ Final Model				

Table 4(a)Telecom1 dataset and (b)features selected

Analogy-X stepwise feature selection procedure shows that the feature *Changes* is the most influential variable (see Table 4b). Further sensitivity analysis shows no abnormal project cases exist in the dataset. The result confirms that the Telecom1 dataset can be safely used for analogy.

Case	Actual Effort	ANGEL Prediction	Absolute Residual	Analogy-X Prediction	Absolute Residual	Absolute Residual Difference
1	305.22	333.96	28.74	374.34	69.12	-40.38
2	330.29	123.1	207.19	1115.54	785.25	-578.06
3	333.96	305.22	28.74	358.37	24.41	4.33
4	150.4	158.56	8.16	117.87	32.53	-24.37
5	544.61	374.34	170.27	123.1	421.51	-251.24
6	117.87	97.45	20.42	158.56	40.69	-20.27
7	1115.54	123.1	992.44	330.29	785.25	207.19
8	158.56	97.45	61.11	168.12	9.56	51.55
9	573.71	168.12	405.59	333.96	239.75	165.84
10	276.95	305.22	28.27	374.34	97.39	-69.12
11	97.45	138.215	40.765	168.12	70.67	-29.905
12	374.34	544.61	170.27	305.22	69.12	101.15
13	168.12	117.87	50.25	158.56	9.56	40.69
14	358.37	333.96	24.41	333.96	24.41	0
15	123.1	330.29	207.19	544.61	421.51	-214.32
16	23.54	34.25	10.71	34.25	10.71	0
17	34.25	23.54	10.71	23.54	10.71	0
18	31.8	34.25	2.45	34.25	2.45	0
Mean	**284.338**	**202.416**	**137.093**	**280.944**	**173.588**	**-36.4953**

Table 5 Predictions using ANGEL and Analogy-X

Validation of analogy prediction based on the feature *Changes* gives a *MMRE* = 0.603; in contrast, ANGEL's brute force based on the feature *Files* gives a *MMRE* = 0.4. The absolute prediction residuals of ANGEL and Analogy-X are displayed in Table 5. We also use a paired *t*-test to examine whether using the feature *File* would produce estimates superior to using the feature *Changes*. The paired *t*-test shows that the prediction produced by either of the two features will produce very similar estimates, where *t*-value = 0.8917 and *p*-value = 0.1925. This indicates that the differences are not statistically significant.

5. KEMERER DATASET

There are 15 completed software projects in the Kemerer dataset. Table 6a shows the project features.

Independent Variable	Description	Type	Mantel-R	p-value
Language	Programming Language	Categorical	See Table 6b	
Hardware	Hardware Resource	Categorical		
KLOC	Kilo Lines of Code	Numerical	0.5520	0.001
RAWFP	Raw Function Points	Numerical	0.6296	0.001
AdjFP	Adjusted Function Points	Numerical	0.6534	0.001

Iteration	Feature subset	Mantel R	p-value	Jackknife R	UCL
1	AdjFP	0.6534	0.001	0.6401	0.8274
1a	AdjFP + Language	0.6653	0.001	0.6483	0.8621
1b	AdjFP + Hardware	0.7954	0.001	0.7370	0.9999

Table 6 (a)Kemerer dataset and (b)features selected

The Kemerer dataset has two categorical variables, therefore the two-stage procedure described in [2] is applied. Stepwise feature subset selection shows that *AdjFP* is the most useful variable. Categorical variable analysis in stage 2 (see [2]) shows the features *Language* and *Hardware* have improved the overall correlation. However the improvement is not significantly better than using *AdjFP* itself, as their Mantel *R* value is less than the upper confidence limit (*UCL* = 0.8274 at 95%) of *AdjFP* itself. Thus only *AdjFP* should be used for the prediction of software effort required in this case. Sensitivity analysis shows no abnormal project cases, indicating the dataset is stable, free of abnormal cases and can be safely used for analogy.

Using ANGEL's brute force feature selection found that the features *KLOC* and *Language* are the optimum feature subset, which gives a *MMRE* of 0.638. In contrast, feature *AdjFP* alone produces a *MMRE* of 0.681.

Based on the estimates from ANGEL and Analogy-X in Table 7 and the paired *t*-test on absolute residuals of the two approaches, show the prediction performance of using either method is not significantly different (statistically), where *t*-value = 1.0964 and *p*-value = 0.2914. The implication is that using either of the subsets derived from ANGEL or Analogy-X will produce similar estimates.

Case	Actual Effort	ANGEL Prediction	Absolute Residual	Analogy-X Prediction	Absolute Residual	Absolute Residual Difference
1	287	258.7	28.3	336.3	308	-279.7
2	82.5	72	10.5	84	73.5	-63
3	1107.31	336.3	771.01	258.7	512.31	258.7
4	86.9	130.3	43.4	230.7	187.3	-143.9
5	336.3	1107.31	771.01	246.9	524.11	246.9
6	84	82.5	1.5	82.5	81	-79.5
7	23.2	82.5	59.3	72	12.7	46.6
8	130.3	86.9	43.4	69.9	26.5	16.9
9	116	258.7	142.7	258.7	116	26.7
10	72	82.5	10.5	23.2	12.7	-2.2
11	258.7	287	28.3	116	87.7	-59.4
12	230.7	157	73.7	86.9	13.2	60.5
13	157	246.9	89.9	86.9	3	86.9
14	246.9	157	89.9	336.3	246.4	-156.5
15	69.9	84	14.1	130.3	116.2	-102.1
Mean	**219.25**	**228.64**	**145.17**	**161.29**	**154.71**	**-9.54**

Table 7 Predictions using ANGEL and Analogy-X

6. DISCUSSION AND CONCLUSION

This study conducted three empirical experiments with the Albrecht, Telecom1 and Kemerer datasets to examine the true efficacy of Analogy-X. At first, it appears that ANGEL is superior in prediction accuracy, but given the problematic *MMRE* metric[4] and further investigation using standard paired *t*-test, there is evidence that both approaches produce similar results. The findings of this study are also consistent with previous studies [2, 3] Nevertheless we must stress that Analogy-X is not a predictive model or providing a means of improving prediction accuracy. It is a data preprocessing procedure to systematically assess the quality of the dataset being used prior to the application of analogy. More importantly, Analogy-X is able to identify relevant features and abnormal project cases based on inferential statistics. This work presents a significant contribution to the knowledge of software effort estimation and other related machine learning techniques.

7. ACKNOWLEDGEMENTS

NICTA is funded through the Australian Government's Backing Australia's Ability Initiative, in part through the Australian Research Council.

8. REFERENCES

[1] M. J. Shepperd and C. Schofield, "Estimating Software Project Effort Using Analogies," *IEEE Transactions on Software Engineering,* vol. 23, pp. 736-743, 1997.

[2] J. W. Keung, B. A. Kitchenham, and D. R. Jeffery, "Analogy-X: Providing Statistical Inferences to Analogy-based Software Cost Estimation," *IEEE Transactions on Software Engineering,* vol. 34, no. 4, 2008.

[3] J. W. Keung and B. A. Kitchenham, "Experiments with Analogy-X for Software Cost Estimation," in *Australian Software Engineering Conference,* Perth, 2008.

[4] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, "A Simulation Study of the Model Evaluation Criterion MMRE," *IEEE Transactions on Software Engineering,* vol. 29, pp. 985-995, 2003.