International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.08, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Selecting a Standard Set of Attributes for Cost Estimation of Software Projects

Arvind Sharma
Assistant Professor
PG Deptt.Of CS&IT,DAV COLLEGE,AMRITSAR

A*bstract*— **the aim of the software engineering is to enhance projects that produce the needed results within limited schedule and budget. So that, software effort estimation becomes a valuable manner since it limits the problems of overestimate and underestimate for the software.**

Software cost estimation is the process of predicting the effort required to develop a software system. There are many estimation models over the last decade, and in this paper, we use six public cost estimation data sets that we obtained from promise repository. We perform regression analysis over these data sets and perform a feature selection in order to get the most effective attribute to the effort. Finally, we analyze and compare the results obtained from each data set to build a framework for the standard set of metrics that we suggest each cost estimation data set must contain.

Keywords- cost estimation, feature selection;ols regression, metrics

I. Introduction

In recent years, software has become the most expensive component of computer system projects; for that, software cost estimation becomes critical and significant to both developers and customers by predicting the effort required to improve a software system. It is essential for both developers and customers to get accurate software cost estimation by accurately estimating the new project cost. Project managers can provide the customers with an accurate deadline for their projects and debate some of the contract negotiation's issues. So, customers can expect actual development costs to be in line with estimated cost. As well as, these estimations can be used by developers to generate reports and proposals in order to determine what resources are needed to commit to the project. These resources will be used as a result of the prioritizing development projects with respect to an overall business plan. Accurate software cost estimation makes project management easier to be managed and controlled as resources are better matched to the real needs [9].

Many researchers in software engineering field have in depth studied how to predict the software project cost which is important for the project managers and software development organizations. Cost estimation, or what is called "effort prediction" is the process of estimating the cost of the software system development. This estimation can generally be estimated through three methods: experts' judgment, algorithmic model and analogy- based method.

Several techniques have been proposed in the past decades in order to make an accurate cost estimation for the projects and then avoid the overruns in the budge and increase the organization efficiency as a result of improving software investment analysis [1]. A key factor in selecting cost estimation model is the accuracy of its metrics since these models depend on their metrics which act as an input to the model. Metric can be defined as a "quantitave measure of the degree to which a system, component, or process possess a given attribute in order to produce a reliable assessment of these attributes in the real problem"[1].

There are too many data sets in the software cost estimation area with different attributes. That makes it difficult and annoying to a person who wants to analyze and correlate the data set in order to use the best and the strongest one in his analysis, and make cost estimation for their projects. So that, there is a necessary need for a valuable standard model. This suggestion can be constructed by obtaining the similarities among those available datasets. This can be applied by first understanding the different attributes for each dataset and finding the effective set of metrics which directly affects effort estimation using regression. After that, we make a comparison among different datasets in order to build a framework which consists of a common set of metrics which serves the user with best prediction, and improves the accuracy estimate of effort required to build a software system.

II. Background

A- Cost estimation

Projects managers need to determine the cost of their ordered projects in order to manage their budgets and determine the deadline of the projects as well as improving the overall quality of any new projects. So, project managers need an effective model which helps them in accurate cost estimation according to their current project requirements.

Heemstra [6] discussed many questions related to software cost estimation; those questions are emphases on the causes for the excesses of the budgets and ruled periods. He explained the preconditions for the estimation and the way for this process. The benefit the software project management can obtain from the used models was explained by viewing the strengths and weaknesses of cost estimation models.

B- Cost estimation methods

Cost estimation methods can be grouped under three categories: expert judgment, analogy- based estimation [10] and algorithmic estimation [9]. In this paper, we are interested in Algorithmic estimation, where the algorithmic model estimates the software cost through some formulas, depending mainly on the size of the project which is measured in terms of Function Point, Object Point and Line of Code (LOC).

In addition to the size of the project, there are several variables participated in the algorithmic model function such as:

Effort = function (var_1, var_2, var_3…var_n)

Where effort is a cost estimation measure that is usually measured by (person-month), function refers to the function form, and (var_1, var_2, var_3…var_n) refers to the cost factors.

In the COCOMO 2 model, Boehm proposed a set of cost factors which are classified into four groups as:-

- Product factor which refers to the software product features like product complexity, database size used…etc.
- Computer factor which refers to the computer characteristics.
- Personal factor which refers to the development staff capability.
- Project factor which and the project environment refers to the project work process.

Through the algorithmic model, some mathematical formula forms can be used such as [9]:

* Linear models.

* Multiplicative models: They refer to the form:

$$(\text{Effort} = a_0 \prod a_i^{var_i}). \quad (1)$$

Where var_i refers to the project factor, a_0, a_1 … are the coefficients extracted from experimental calculations.

* Power function model: They refer to the form $(\text{effort} = a * s^b)$. (2)

Where s is the software size usually measured by lines of code (LOC), a and b are the coefficients extracted from experimental calculations of the data set.

COCOMO (Constructive Cost Model), proposed by Boehm, is an example of the algorithmic model, which is a software package that helps assisting projects managers in planning a cost estimation of a software development project through an interactive interface. Furthermore, COCOMO is a highly subjective to the users input variables by noting the equations coefficients.

Korte and Port [8] implemented a standard and easily-analyzed statistical methods –standard error and bootstrapping –to several COCOMO 1 model research findings. The primary focus is on the confidence obtained from the findings that are experimentally based on estimators for error distribution parameters. As a result, this will reduce the contradictory and the lack of confidence in several published cost estimation research results based on Precisely MMRE and PRED comparisons such as model selection.

C - Feature subset selection

Feature selection which is referred to as subset selection is a pre- processing procedure used in machine learning where a subset of the features obtained from the data is chosen for implementation of a learning algorithm. The best subset includes the minimum number of diminutions which increase the accuracy by eliminating the inefficient dimensions [11].

Das and Kempe [5] demonstrated the problem of selecting a subset of k random variables to perceive that will produce the best linear prediction of different important variable, taking into account the pair wise correlations between the other variables and the predictor variable.

Azzeh, Neagu and Cowling [2] checked the influence of using feature subset selection algorithms in enhancing the accuracy of analogy software effort estimation models. They validated their works using two established data sets (ISBSG and Desharnais) use MMRE as evaluation criteria for all feature subset selection algorithmsl. They concluded that the employment of a fuzzy feature subset selection algorithm in analogy software effort estimation can give a valuable result.

Menzies, Port and Boehm [4] discovered that COCOMO's estimates can be enhanced by using WRAPPER which is a feature subset selection method improved by the data mining industry. The results showed that the features subset selection always enhance the PRED(30) values without rising variance when applying on different data sets and as a result this will improve the strength of the COCOMO's prediction.

Kirsopp, Sheppered and Hart [7] explained the employment of using search techniques to aid the enhancement of case –based reasoning (CBR) system used in the software effort prediction. They checked the use of random searching, hill climbing and forward sequential selection (FSS) in order to get the optimal feature subsets that affect the effort prediction. They concluded that using a random search is better than using all the features. However using hill climbing and FSS can give better results than random search. They suggested using some form of heuristic –based initialization that can help in gaining improved results.

III. The Methodology

Software cost estimation can be considered as an empirical process that can be used to calculate the effort and the

IJARSE
ISSN: 2319-8354

development time requirements for the software product, and hence this process will affect the overall managing, developing and scheduling of the software projects. The real problem that the researches may be faced is how to generate useful software cost predictions at an early stage in a project. Public data sets from a real project can be analyzed and evaluated to solve this problem. In this paper, public data sets are analyzed in order to obtain the standard set of metrics that mainly affect the effort.

Our methodology is displayed in Figure 1. It starts with cost estimation data sets collection that is listed in PROMISE repository (http://promisedata.org). We select six data sets that have a somewhat similar attributes. In the second phase, we analyze the different public datasets manually in order to specify the attributes for each data set and the description for each attribute. Then, we perform a regression analysis over different public data sets in order to obtain the best selected attributes that affect cost estimation process. Finally, we make a comparison among the selected attributes using regression from different data sets to build a framework for a standard set of metrics. Phase 3 and phase 4 will be discussed in the following sections.

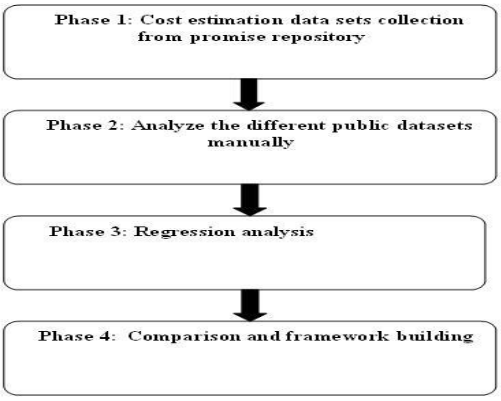


Figure 1 Methodology

IV. Feature selection using OLS regression analysis

Regression methods are widespread in the last decade. The most usually used regressions method is the ordinary least square (OLS) regression which has also been criticized since it has restrictions [3]. OLS regression is one of the most popularly used models for cost estimation. Linear regression has the following form:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_n X_n + e \qquad (3)$$

Where $\hat{Y}$ denotes the dependant variable (effort for the project) and X_i refers to as independent variables (features or effort drivers), and b_i is called regression coefficient, a is referred as the intercept, and the error term e is a random noise with a normal distribution.

The OLS method works in the way that it estimates the response coefficients and the intercept parameter by minimizing the least square error term R^2_i where R_i is the estimating the response coefficients and the intercept parameter by minimizing the difference between the observed response and the model predicted response for the i^{th} observations.

We use a Minitab statistical tool in order to perform the ordinary least square regression (OLS) over six public dataset: cocomo81, cocomo_sdr, cocomonasa, kemerer, Maxwell and nasa93. We first perform a regression analysis for all the attributes in each data set in the effort to obtain the possible effective attributes. This may be inferred when the value for p or the significance level is less than 0.05. The p value of 5% means that there is a 5% chance that the relationship emerged randomly and a 95% chance that the relationship is real. We also take into account the R square and the adjusted R square (R square (adj)) for each model that measures how the model fits the data. R square is called the coefficient of determination or the percentage of variance explained.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \qquad (4)$$

The adjusted R square is particularly useful in the feature selection stage of the model building which gives us the best feature subsets.

Figure 2 shows the different building models with different attributes from performing the regression analysis for all the attributes of the cocomo81 data set. From this model, the effective attributes are the attributes that have the largest R square (adj) value.

For cocomo81 data set, when we perform the best subset regression, we obtain the following figure that shows the models building with various numbers of attributes; then, we select the best model which has the largest R square (adj) value.

Best Subsets Regression: actual versus rely; data; ...

Response is actual

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	rely	data	cplx	time	stor	virt	turn	acap	aexp	pcap	vexp	lexp	modp	tool	sced	loc
1	43.2	42.3	12.4	1384.1																X
1	20.2	18.9	41.3	1640.6		X														
2	48.5	46.8	7.7	1328.5													X			X
2	48.1	46.4	8.3	1334.2		X														X
3	54.1	51.8	2.7	1264.9	X												X			X
3	52.8	50.4	4.3	1282.7				X									X			X
4	58.1	55.2	-0.4	1218.7	X	X											X			X
4	57.8	54.8	0.1	1224.0		X	X										X			X
5	59.4	55.9	-0.0	1210.0	X	X						X					X			X
5	59.4	55.8	0.0	1210.4	X	X	X										X			X
6	61.1	57.0	-0.1	1195.1	X	X						X	X				X			X
6	60.5	56.2	0.7	1205.2	X	X	X		X								X			X
7	61.6	56.7	1.3	1199.1	X	X	X	X	X								X			X
7	61.5	56.6	1.4	1199.9	X	X			X			X	X				X			X
8	62.3	56.7	2.4	1198.8	X	X		X	X			X	X				X			X
8	62.1	56.5	2.6	1201.2	X	X	X		X			X	X				X			X
9	62.9	56.6	3.6	1199.7	X	X	X	X	X			X	X				X			X
9	62.4	56.0	4.3	1208.3	X	X		X	X			X	X				X	X		X
10	63.0	55.9	5.5	1210.2	X	X	X	X	X			X	X			X	X			X
10	63.0	55.9	5.5	1210.3	X	X	X	X	X	X		X	X				X			X
11	63.2	55.3	7.2	1218.1	X	X	X	X	X	X		X	X		X		X			X
11	63.2	55.3	7.3	1218.5	X	X	X	X	X	X		X	X			X	X			X
12	63.3	54.5	9.1	1229.0	X	X	X	X	X	X	X	X	X		X		X			X
12	63.3	54.5	9.2	1229.4	X	X	X	X	X	X		X	X		X	X	X			X
13	63.3	53.6	11.1	1240.5	X	X	X	X	X	X	X	X	X		X	X	X			X
13	63.3	53.6	11.1	1240.8	X	X	X	X	X	X	X	X	X		X		X	X		X
14	63.4	52.7	13.0	1252.7	X	X	X	X	X	X	X	X	X		X	X	X	X		X
14	63.4	52.7	13.1	1253.1	X	X	X	X	X	X	X	X	X	X	X	X	X			X
15	63.4	51.7	15.0	1265.6	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X
15	63.4	51.7	15.0	1265.9	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.08, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Figure 2 Models building with different values of R square (adj)-cocomo81

The first model in Figure 2, we have one of the effective attributes which is loc attribute with 43.2 and 42.3 for R square and R square (adj), respectively. The second model also, has one effective attribute which is data with R square equals to 20.3 and Square (adj) equals 18.9. So that, the best model which has the largest R square (adj) is 57 having six effective attributes (rely, data, acap, aexp, modp and loc). After that, we perform a regression method to the selected six attributes we got the following regression equation.

Effort= -12170+1980 rely +5670 data -2378 acap + 2218 aexp + 4694 modp+5.76 loc. (5)

The over all p value in ANOVA table can be used in order to evaluate the model which measures the similarity in which the model as a whole describes a relationship that emerged at a random, rather than a real relationship. Table 1 shows the ANOVA distribution for the model with an over all p value less than 0.05. This indicates that the regression model contains an effective set of attributes that affect the response attribute.

Table 1 ANOVA for the best model

Source	DF	SS	MS	F	P
Regression	6	125739069	20956511	14.67	0.0
Residual Error	56	79986991	1428339		
Total	62	20572605			

For cocomo_sdr data set, the best model cannot be obtained in one step since the number of the coefficient is greater than or equal to the number of observations. So that, we perform a regression in several steps in to obtain the best effective attributes with regard to the R square (adj) value. So, we calculate the correlation among in an effort to determine the best set of attributes which can form the model of regression. The obtained attributes are: PMAT, FLEX, TIME, PVOL, PREC, RESL and LOC are effective attributes with the following equation:

Effort = 60.3-2.52 PMAT +6.86 FLEX -27.3 TIME- 34.9 PVOL – 4.12 PREC + 2.50 RESL +0.000099 LOC (6)

Best Subsets Regression: ACTUAL EFFORT versus FLEX; DOCU; ...

Response is ACTUAL EFFORT

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	FLEX	DOCU	PMAT	RELY	TIME	PVOL	LOC	PREC	RESL
1	61.1	57.2	94.4	4.4769							X		
1	24.7	17.1	190.1	6.2264	X								
2	69.0	62.1	75.6	4.2125			X				X		
2	67.8	60.6	78.7	4.2916							X	X	
3	74.5	65.0	63.0	4.0475			X				X	X	
3	72.6	62.4	67.9	4.1946		X	X				X		
4	82.9	73.1	43.1	3.5501	X		X				X	X	
4	82.4	72.4	44.2	3.5956	X				X		X	X	
5	89.5	80.7	27.7	3.0055	X	X	X	X			X		
5	87.2	76.5	33.7	3.3142	X		X				X	X	X
6	92.3	83.0	22.3	2.8168	X		X			X	X	X	X
6	92.1	82.5	22.9	2.8597	X		X		X	X		X	X
7	99.0	97.4	6.5	1.1067	X		X		X	X	X	X	X
7	93.8	83.0	20.3	2.8200	X	X	X		X		X	X	X
8	99.1	96.7	8.4	1.2425	X		X	X	X	X	X	X	X
8	99.1	96.6	8.4	1.2544	X	X	X		X	X	X	X	X
9	99.2	95.8	10.0	1.3990	X	X	X	X	X	X	X	X	X

Figure 3 Models building with different values of R square (adj)-cocomo_sdr

The first model we have consist of one attribute which is loc and second has also one attribute which is FLEX but the best model have 7 attributes as discussed previously with R square =99.0% and R square (adj) = 97.4 %

For Kemerer data set, the best model is displayed in Figure 4 that shows the selected attributes: hardware, duration and adjfp with R square =71.9% and R square (adj) =64.2%.

After performing the regression method to the selected three attributes, we got the following regression equation:

Effort = - 259 + 63.1 Hardware - 6.83 Duration + 0.429 AdjFP (7)

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	Language	Hardware	Duration	KSLOC	AdjFP	RAWFP
1	58.5	55.3	1.4	175.96					X	
1	57.1	53.8	1.8	178.85						X
2	68.9	63.7	0.3	158.43		X			X	
2	67.0	61.5	0.8	163.28		X				X
3	71.9	64.2	1.4	157.29		X	X		X	
3	69.8	61.6	2.0	163.02		X		X	X	
4	72.5	61.5	3.2	163.29	X	X	X		X	
4	72.4	61.3	3.2	163.57		X	X	X	X	
5	73.0	58.0	5.1	170.55		X	X	X	X	X
5	72.6	57.5	5.2	171.59	X	X	X	X	X	
6	73.2	53.0	7.0	180.27	X	X	X	X	X	X

Figure 4 models building with different values of R square (adj) - kemerer

For the Maxwell data set, we cannot obtain the best model with only one step because the attributes are highly correlated. For that reason, we calculate the correlation among the attributes and the effort in order to determine the effective set of attributes. Then, we obtain the best model with high values of R square (adj) that equal to 81.2% and R square equal 83.6% the attributes are : Source, Telonuse, Development environment adequacy (T02), Staff application Knowledge (T13), Staff tool skills (T14), Staff team skills (T15), duration and size can be effective attributes with the following equation for the regression.

Effort = - 10722 + 6890 Source - 3854 Telonuse - 1104 T02 - 1006 T13 - 1454 T14 + 2060 T15 + 302 Duration + 9.06 Size (8)

Best Subsets Regression: Effort versus Source; Telonuse; ...

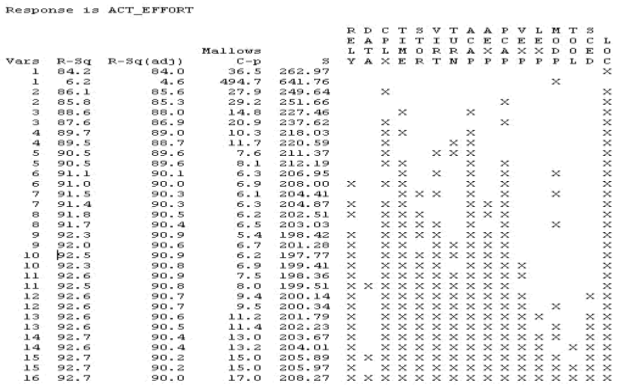
Response is Effort

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	Source	Telonuse	T02	T13	T14	T15	Duration	Size
1	70.9	70.4	36.4	5714.7								X
1	43.1	42.2	126.1	7982.5							X	
2	77.4	76.6	17.2	5075.4							X	X
2	74.3	73.5	27.1	5407.6					X			X
3	79.4	78.3	12.7	4886.5	X						X	X
3	79.1	78.0	13.8	4928.1					X		X	X
4	80.7	79.3	10.6	4774.6	X				X		X	X
4	80.1	78.7	12.6	4851.2	X	X					X	X
5	81.6	79.9	9.6	4702.3	X	X			X		X	X
5	81.3	79.7	10.4	4733.1	X				X	X	X	X
6	82.6	80.6	8.5	4618.9	X	X			X	X	X	X
6	82.1	80.2	9.9	4674.7	X	X	X		X		X	X
7	83.2	81.0	8.6	4579.9	X	X		X	X	X	X	X
7	83.0	80.8	9.1	4602.8	X	X	X		X	X	X	X
8	83.6	81.2	9.0	4556.7	X	X	X	X	X	X	X	X

Figure 5 Models building with different values of R square

(adj) – Maxwell

For cocomonasa data set, the model is displayed in Figure 6 that shows the selected attributes with R square (adj) value



Response is ACT_EFFORT

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	RELY	DATA	CPLX	TIME	STOR	VIRT	TURN	ACAP	AEXP	PCAP	VEXP	LEXP	MODP	TOOL	SCED	LOC
1	84.2	84.0	36.5	262.97																X
1	6.2	4.6	494.7	641.76													X			
2	86.1	85.6	27.9	249.64			X													X
2	85.8	85.3	29.2	251.66										X						X
3	88.6	88.0	14.8	227.46				X				X								X
3	87.6	86.9	20.9	237.62			X							X						X
4	89.7	89.0	10.3	218.03			X	X				X								X
4	89.5	88.7	11.7	220.59			X				X	X								X
5	90.5	89.6	7.6	211.37			X			X	X	X								X
5	90.5	89.6	8.1	212.19			X	X				X		X						X
6	91.1	90.1	6.3	206.95				X		X		X		X			X			X
6	91.0	90.0	6.9	208.00	X		X	X				X		X						X
7	91.5	90.3	6.1	204.41				X	X	X		X		X			X			X
7	91.4	90.3	6.3	204.87	X		X	X				X	X	X						X
8	91.8	90.5	6.2	202.51	X		X	X	X			X	X	X						X
8	91.7	90.4	6.5	203.03			X	X	X	X		X		X			X			X
9	92.3	90.9	5.4	198.42	X		X	X	X	X		X	X	X						X
9	92.0	90.6	6.7	201.28	X		X	X		X	X	X	X	X						X
10	92.5	90.9	6.2	197.77	X		X	X	X	X	X	X	X	X						X
10	92.3	90.8	6.9	199.41	X		X	X	X	X		X	X	X	X					X
11	92.6	90.9	7.5	198.36	X		X	X	X	X	X	X	X	X	X					X
11	92.5	90.8	8.0	199.51	X	X	X	X	X	X	X	X	X	X						X
12	92.6	90.7	9.4	200.14	X		X	X	X	X	X	X	X	X	X				X	X
12	92.6	90.7	9.5	200.34	X		X	X	X	X	X	X	X	X	X		X			X
13	92.6	90.6	11.2	201.79	X		X	X	X	X	X	X	X	X	X	X			X	X
13	92.6	90.5	11.4	202.23	X		X	X	X	X	X	X	X	X	X		X		X	X
14	92.7	90.4	13.0	203.67	X		X	X	X	X	X	X	X	X	X	X	X		X	X
14	92.6	90.4	13.2	204.01	X		X	X	X	X	X	X	X	X	X	X		X	X	X
15	92.7	90.2	15.0	205.89	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X
15	92.7	90.2	15.0	205.97	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X
16	92.7	90.0	17.0	208.27	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

is 90.9% and R square value equal to 92.3%.

Figure 6 Models building with different values of R square (adj) – cocomonasa

After applying the regression method, we got the following equation:

Effort = - 2254- 724 rely + 910 cplx + 1261 time -543 stor - 702 virt + 1991 acap -1135 aexp + 950 pcap + 7.05 loc. (9)

As to nasa93, we have the following equation for regression.

Effort = 4771-1174 pr1- 1217 pr2-1615 pr3 – 2219 pr4 -802 pr5 -2368 pr6 -1712 pr7 -625 C1+ 1545 C2– 404 C3+ 108 C4-419 C5+ 159 C6+ 1210 C7– 307 C8+539 C9– 801 C10-438 C11+ 474 C12– 72 C13-363 center + 1242 time +6.24 equivphyskloc – 2164 virt -2030 modp. (10)

It is noteworthy to mention that we perform some operations on the variables that are text in order to make the regression available; here we have the project name and cat2 that are indicator variables. The following table shows the indication of the used attributes.

Table 2 Indicator variables

Attribute	category
Pr1	De
Pr2	erb
Pr3	gal
Pr4	hst
Pr5	slp
Pr6	spl
Pr7	x
C1	Application groubd
C2	avionics
C3	avionicsmonitoring
C4	batchdataprocessing
C5	communications
C6	datacapture
C7	launchprocessing
C8	missionplanning
C9	monitor_control
C10	operatingsystem
C11	realdataprocessing
C12	science
C13	simulation

V. Standard Set of Attributes

Loc appears in different data sets with different names such as loc, ksloc and equivphyskloc. We have some attributes with different names such as T13 staff application knowledge is the same with aexp, and T15 (staff team skills) is the same to team.

Table 3 Resulted attributes from regression analysis

Data set	Attributes
Cocomo81	Rely, data, acap, aexp, modp, loc
Cocomo_sdr	Pmat, flex, time,pvol,prec,resl,loc
cocomonasa	Rely, cplx, time, stor, virt, acap, aexp, pcap, loc
kemerer	Hardware, duration, adjfp
Maxwell	Source,telonuse,T02,T13,T14,T15,duration,size
Nasa93	Program name, cat2,time,virt,modp,equivphyskloc

Table 4 Framework for the standard set of metrics used for effort estimation

Category	Standard attributes
Product Factors	Rely data cplx duration pmat flex prec
Platform Factors	Hardware pvol Time stor virt Source T02 (development environment adequacy)
Personnel Factors	acap aexp pcap modp team T14 (Staff tool skills)
Project Factors	Loc adjfp Size project name Cat2 telonuse resl

VI. Conclusion

Cost estimation plays a significant role in the software development since it attempts to avoid software overestimate or underestimate for the allocated budget. Various cost estimation models exist, but the effectiveness of each depends basically on its constituent attributes. In this paper, we make a feature subset selection over public selected data sets: cocomo81, cocomo_sdr, cocomonasa, Kemmerer, Maxwell and nasa93. We use regression analysis to obtain the best model built using minimum number of attributes. The results from the selected attributes are compared in order to build a framework of the standard set of metrics for effort estimation. This framework consists of attributes that may be classified into four categories: product factors, platform factors, personnel factors and project factors.

References

[1] Ayyildiz, M., Kalipsiz, O., &Yavuz, S. (2006).A Metric-Set and Model Suggestion for Better Software Project Cost Estimation. World academy of science, Engineering and Technology, 23,167-172.

[2] Azzeh, M.Neagu, D. & Cowling. (2008).improving analogy software effort estimation using a fuzzy feature subset selection algorithm, promise 08 proceeding of the 4th international workshop on predictor models in software engineering, 71-78.

[3] Briand, L., Emam, K., Surmann ,D., Wieczorek, I., & Maxwell, K. (1999).an assessment and comparison of common software cost estimation modeling techniques , ICSE '99 proceedings of the 21st international conference on

software engineering,313-322.
[4] Chen, Z.Menzies, T.Port, D. &Boehm, B. (2005).feature subset selection can improve software cost estimation accuracy. PROMISE '05 proceedings of the 2005 workshop on predictor models in software engineering, 1-6.
[5] Das, A.Kempe, D. (2008).Algorithms for subset selection in linear regression. STOC 08 proceedings of the 40th annual ACM symposium on theory of computing, 1-10.
[6] Heemstra, F.J. (1992). Software cost estimation. Information and software technology, volume 34, issue 10, 627-639.
[7] Kirsopp, C.Shepperd, M.Hart, J. (2002).search heuristic, case –based reasoning and software project effort predicrion.GECCO 02 proceedings of the genetic and evolutionary computation conference, 1-8.
[8] Korte, M. & Port, D. (2008) confidence in software cost estimation results based on -MRE and PRED. PROMISE'08 proceedings of the 4th international workshop on predictor models in software engineering, 63-70.
[9] Leung, H., & Fan, Z. (2002). Software Cost Estimation. Software engineering and knowledge engineering, 1-14.
[10] Liyan-Fu. (2009).Improvement and Implementation of Analogy Based Method for Software Project Cost Estimation (doctoral thesis).Retrieved from http:// scholarbank.nus .edu.sg/bitstream/handele /10635/17724/LIYF.pdf.
[11] Sewell, M. (2007). Feature selection. Available on http:// machine - learning. Martinsewell.com.