European Journal of Applied Sciences 8 (4): 257-269, 2016
ISSN 2079-2077

DOI: 10.5829/idosi.ejas.2016.8.4.22533

Empirical Evaluation of Different Machine Learning Methods for Software Services Development Effort Estimation Through Correlation Analysis

[1]Amid Khatibi Bardsiri and [2]Seyyed Mohsen Hashemi

[1]Department of Computer Engineering, Science and Research Branch,
Islamic Azad University, Tehran, Iran
[2]Science and Research Branch, Islamic Azad University, Tehran, Iran

Abstract: The concept of development effort generally means the time or the cost of developing a software service. An essential factor to successfully manage and control a project is the accurate estimation of the development effort and an over and underestimation lead to the loss of project resources. So far, different effort estimation models have been presented in three domains: expert judgment, algorithmic methods and machine learning methods. Recently, several approaches in the last domain, machine learning, have been applied for software service development effort, which had a higher performance in comparison to the other two domains. This paper presents an empirical evaluation of the performance and accuracy of five main machine learning methods using the correlation analysis approach and investigates the effects of feature selection on the estimation accuracy. The evaluations and comparisons are performed using two well-known and real-world datasets, NASA and ISBSG and two artificial datasets, Moderate and Severe. Finally, the obtained results provide a clear illustration of the performance of these machine learning methods and the effects of feature selection on the estimation accuracy.

Key words: Effort estimation · Software service · Machine learning · Empirical evaluation · Correlation analysis

INTRODUCTION

On time and budget determined delivery of service, is one of the main concerns of the most software companies. The necessary effort to develop a software service is among the most important and effective parameters of a project. Since the estimation process should be carried out in initial phases of the project, a reliable method is needed to be able to work with initial and little data [1]. Different methods have been proposed to predict the effort which can be categorized in six groups:Parametric methods such as SEER-SEM, COCOMO [2], Expert judgment such as WBS, Delphi methods [3,4], Learn base models such as ABE [4], Regression methods such as OLS, ROR [5], Dynamic models andHybrid models [6].

The introduction of function point (FP) by Albrecht in 1983, was one of the important events in software measurement which gave the possibility of measuring the first levels of the project and prevented the negative effects of the previous method, LOC [7]. Many changes in software development methodology and progress in estimation methods resulted in development of a new model called COCOMO II by Boehm in 2000 [8]. On the other hand, because of the inability of algorithmic methods in controlling dynamic behavior of software projects and the lack of completeinformation of a project in primary stages, non-algorithmic methods have been presented.

Expert judgment method which was presented in 1963 is an example of these methods [9]. In this method, expert people share their ideas about the estimate value to achieve an agreement. CART, is another method of non-algorithmic method groups which attain the effort value in the leaves of the trees by making a tree and using the previous projects [10]. The most popular non-algorithmic estimation method is ABE method which was presented in 1997 [11]. This method uses comparison of a project with other similar historical cases. The comparison is

Corresponding Author: Amid Khatibi Bardsiri, Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.

based on the features of two projects. Moreover, other smart methods such as neural network, fuzzy rules and different methods of data mining have been used in effort estimation area [6, 12, 13].

In recent years, machine learning techniques have been used extensively in the field of estimating effort and have shown good performance [6, 14-16]. Despite the many improvements, yet are not well defined status of each of these methods and researchers are having difficulty in choosing them. The purpose of this article is the assessment and detailed comparison of different types of these methods.

This paper has been organized in 5 sections. The second section reviews the related works. Sections 3 describemachine learning methods. The empirical evaluation has been presented in section 4 and section 5 includes conclusions.

Related Works: Many techniques have been introduced in the past years to estimate the required effort and cost for developing a software service.These methods have been initiated by simple equations and assumptions and have now achieved complicated techniques [14]. These techniques can be divided into the three general groups below:

Expert Judgment: In this method, which was proposed in the late 1960s [9] and is still widely employed in various software companies, domain experts are asked to give their opinions on the required effort. Various amounts are expressed and, typically, their median is returned as the final required effort. The Delphi method is an example of this class of techniques [17].

Algorithmic Models: These models, which use mathematical relations and equations, seek to discover a relationship between service attributes and the required effort, are usually suitable for specific cases and are adjusted and calibrated depending on the existing conditions. COCOMO, SLIM, SEER-SEM are examples of this type of methods [18].

Machine Learning: These methods look to construct and study algorithms that can learn from datasets, are applied to inputs of the related problem and help in the decision-making. Fuzzy theory, decision tree, ANN and regression are examples of this class of methods [16].

Some of the benefits of machine learning methods include the ability to model complex relationships between dependent and independent variables and also power of learning from historical data.One of the disadvantages of the algorithmic methods, lack of flexibility and the need to calibrate themselves. These methods also do not have the ability to find the complex relationships between variables. Different kinds of regression [6], COCOMO models and COCOMO II [8, 19] are the most famous algorithmic models and ABE [11], CART [10], Expert judgment [9] and Artificial Neural Network [20], Learning and artificial intelligence techniques [21], fuzzy rules and optimization algorithms [22] are the most popular non-algorithmic methods. Figure 1 shows the different types of effort estimation methods and their subsets.

Machine Learning Methods: In this section, briefly be explained 5 different methods of the most important machine learning and continues to evaluate and compare these methods. It is important to note that nature of these models is different with each other completely [6, 14].

SWR and MLR: Regression methods are among the oldest estimation methods and try to fit a function to a set of data. The dataset includes a dependent variable E and several independent variables X_i and the linear Equation 1 is considered for the data [23-25]:

$$Y = B_1X_1 + B_2X_2 + B_2X_2 + ... + B_nX_n + b \quad (1)$$

In this equation, B is the slope of the line and b the value of the intercept, which can obviously be obtained by adding the one's column to the X vector In regression models, the purpose is to find the B and b coefficients in such a way that error is minimized. MLR (Multiple Linear Regression) and SWR (StepWise Regression) are examples regression models [26].

CART: The purpose in CART(Classification And Regression Trees) is to build a structured decision tree for classifying the set of instances in the dataset. The partition criterion is the simple testing of the features of the instances and the tree is built recursively using simple if-then rules [10]. Each instance, depending on the values of its features, moves on the tree and reaches a specific leaf (which, here, is the amount of effort). This model was used in some of the previous studies [23, 24, 27-29].

Analogy Based Estimation (ABE): The ABE model was introduced in 1997 by Schofield and Shepperd as an alternative for the algorithmic techniques [11]. In this model, the effort value is obtained bycomparison of one service with similar and previously completed services

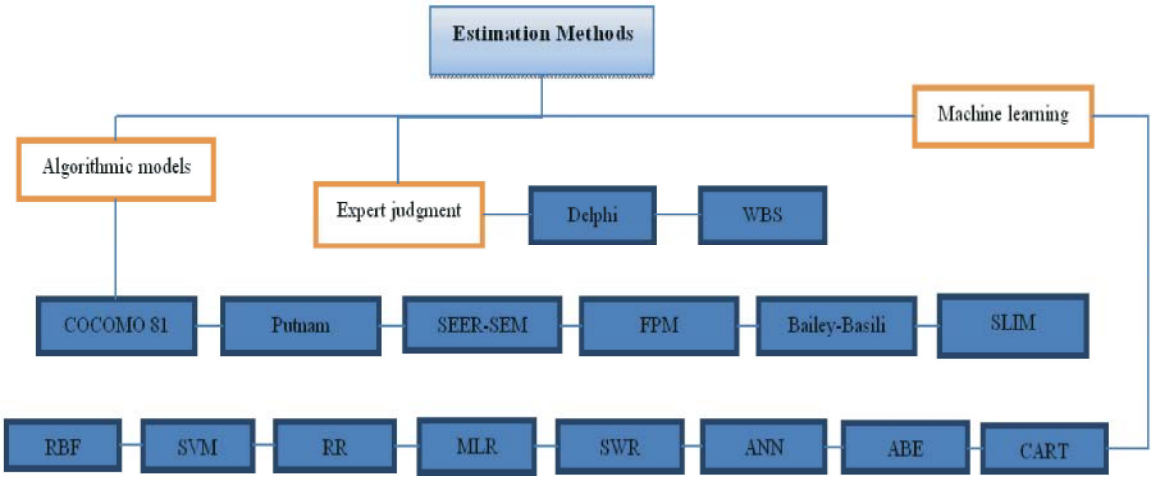

Fig. 1: Various types of effort estimationmethods

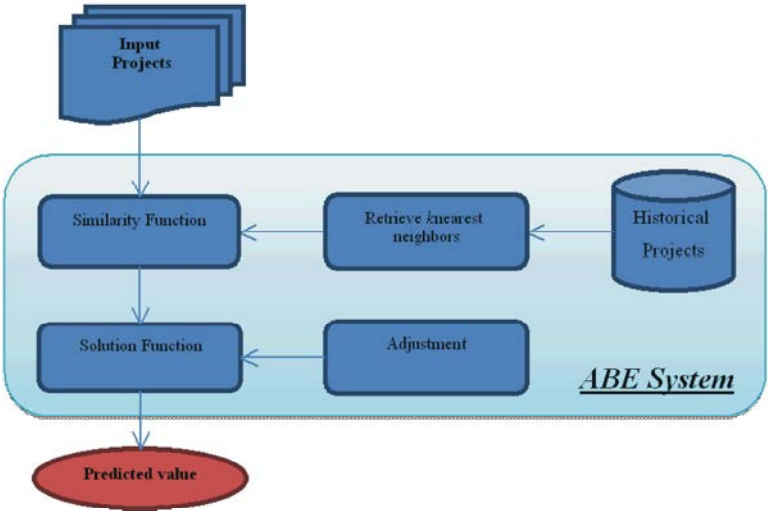

Fig. 2: Analogy Based Estimation diagram

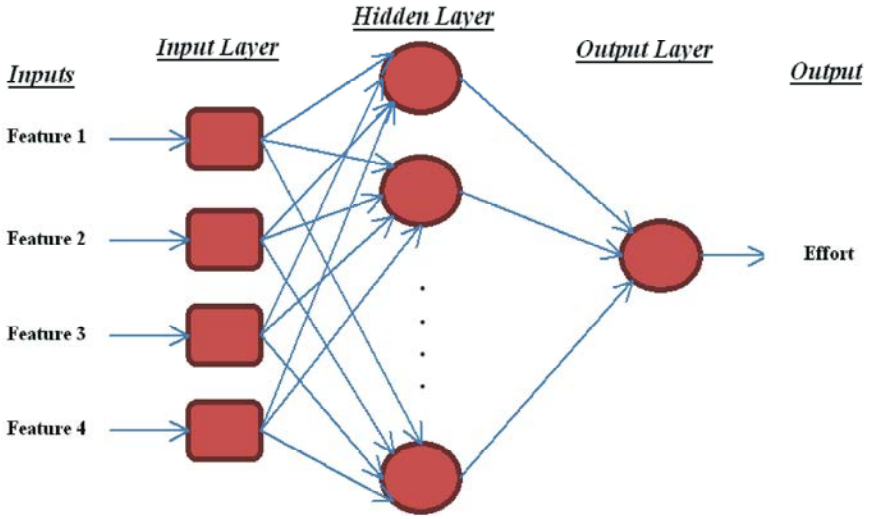

Fig. 3: Simple neural network architecture

(historical cases). In fact, by using Similarity Function, ABE finds the similarities of one service with the similar services (based on the service features) and after selecting some appropriate services (called analogies and shown by KNN parameter) the final solution will be found using Solution Function.The graphic scheme of ABE method is given in Figure 2.

ANN: The neural network is a nonlinear model that imitates the function of human brain and has frequently been used for estimating effort [13]. The neural network consists of a set of neurons in several layers that transport incoming information on their outgoing connections to other units through weighting and by using a suitable transfer function.To generate the output, the inputs take the weight and bias of each neuron and the transfer function processes the inputs of each neuron [30, 31]. The simple neural networkmodel presented in Figure 3.

Empirical Evaluation: This section explains simulation and comparison of the described methods. The simulation was performed with the help of Matlab powerful software and the objective was to compare accuracy estimates of machine learning methods.

Datasets: In order to create and evaluate the associated estimators, fourdatasets were used, two real datasets: ISBSG and NASA and two artificial datasets: Moderate and Severe.The followings are descriptions of each dataset.

ISBSG Dataset: ISBSG is a great company located in Australia [32]. This paper uses the existing data on 11th release of ISBSG dataset which includes partial information of 5052 software projects. This repository, which uses 109 features for each project, has collected its information from 24 different countries.An appropriate filter is required for selecting an applied and reliable subset of ISBSG projects. In the first step, the project with quality rates other than A and B were removed; therefore there was no doubt in the accuracy of the data. Then the projects were filtered by someresource level other thandevelopment, so that the learning effort and alike are not considered in them(resource level $\leq$ 1). Finally, the projects that measurement metric of their sizes were other than IFPUG were removed.In the end, by following the above-mentioned filters, 66 software services were obtained and the research was continued on them.Among all the present features, six important ones [Input count, Output count, Enquirycount, File count, Interface count andAdjusted function point] were selected that influenced the development effort [Normalized effort in hours]. Statistical information of ISBSG dataset is given in Table 1.

Nasa Dataset: The second employed dataset consists of NASA`s known projects, whose statistical information is presented in Table 2. This dataset was first introduced by Bailey and Basili [33] and later used extensively in different studies [6, 34]. NASA dataset contains 18 observations that belong to its different software projects. This dataset includes two independent variables, *DL* (Development Line) and *M* (Methodology) and a dependent variable, *E* (Effort). Variable *DL* refers to the number of development lines in the application, including comments. Variable *M* is a combinatory measure of software development methodologies and variable *E* indicates software management and programming efforts, measured in man-man units.

Artificial Datasets: Unfortunately, real datasets are often old and small with a large number of outliers. Therefore, we are forced to use artificial data to test models. Pickard *et al*. [35] introduced a method for generating simulated data. Equation 2 shows the basis of their method [22, 36].

$$y = 1000 + 6x_1sk + 3x_2sk + 2x_3sk + e_{het} \quad (2)$$

In this equation, x_1sk, x_2sk and x_3sk are independent variables obtained from the gamma random distribution of the variables x_1, x_2 and x_3 (with a mean of 4 and variance of 8). The next variable is relative error calculated from the following equation.

$$e_{het} = c \times e \times x_1sk \quad (3)$$

In this equation, e is the random error resulting from normally distributed random variable with a mean of zero and a variance of 1 and, finally, c is a constant. An artificial dataset has the three main features of variance, skewness and outlier; and two different types of datasets are obtained depending on the values of these three features. Values of the outliers are obtained through multiplication and division of a percentage of the data by specific constant values. Table 3 shows the adjustments made in the two artificial datasets and the values related to them.

Table 1: Descriptionof ISBSG dataset

Variable	Minimum	Maximum	Mean	Median	Std
InpCont	3	1185	169	95	199
OutCont	10	698	143	67	165
EnqCont	3	653	150	116	137
FileCont	7	384	129	108	97
IntCont	5	497	76	43	95
AFP	107	2245	672	507	534
NorEffort	562	60826	6860	4899	8406

Table 2: Descriptionof NASA dataset

Variable	Minimum	Maximum	Mean	Median	Std
DL	2.1	100.8	33.58	17.15	31.67
M	19	35	27.77	18.5	5.23
Effort	5	138.3	49.47	26.2	44.43

Table 3: Descriptionof artificial data sets parameters

Dataset	Relative error constant	Outlier constant	Outliers percentage
Moderate	0.1	2	%5
Severe	6	6	%10

Figures 4 and 5 indicate scatter plots of the 100 data items produced by the datasets Severe and Moderate, respectively. The presence of scattered and deviant values resulting from random distribution causes estimates made by artificial datasets also to be difficult and accompanied by error, so that they can be used in constructing and evaluating models.

Evaluation Criteria: This study aims to compare the accuracy of different approaches and thus two well-known and accepted measures, PRED and MMRE, are employed. Moreover, the statistical method LOOCV (Leave One Out Cross Validation)is used to validate the results. In this method, each time a project is selected as a test case and the remaining projects are used as training data; this process is repeated for the total number of projects. This approach is the only reliable method of validating the obtained results [37]. The use of this technique will increase the validity of the results and the probability that there will be a larger number of random selections. A basic question and, in fact, the most important parameter in any evaluating and estimation method, is its degree of accuracy: how far the estimated value is from the actual one. Equation 4 shows the relative error (RE) for evaluating the efficiency of a method. In this equation, E is the amount of the actual effort and E' the expected, or estimated, amount [11].

$$RE = \frac{E' - E}{E} \quad (4)$$

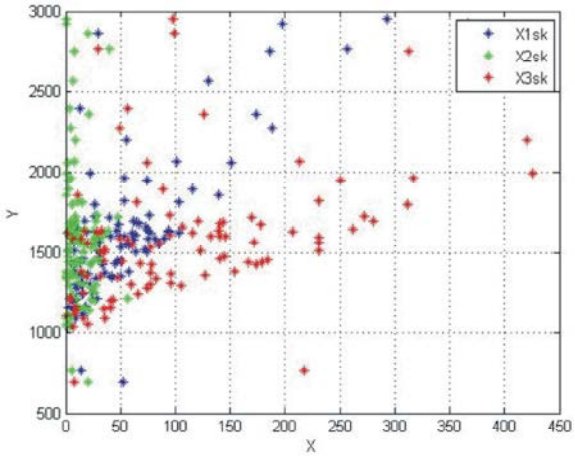

Fig. 4: Y versus X values of artificialModerate dataset

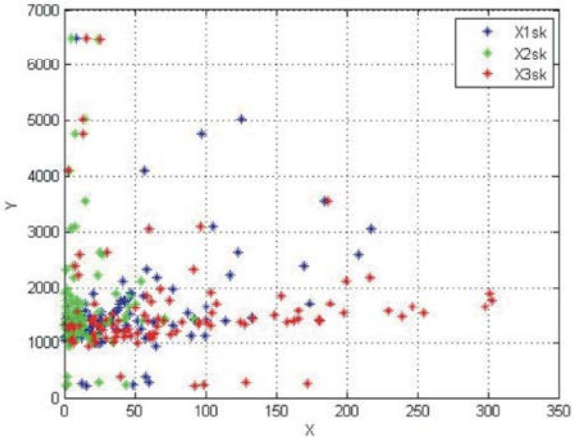

Fig. 5: Y versus $X$of artificial Severe dataset

The *MRE* parameter is an important and commonly used criterion in estimation and its value for a service is shown in Equation5. In fact, *MRE* is the absolute error in estimating project and the lower it is, the more efficient the related method.

$$MRE = \frac{|E' - E|}{E} \quad (5)$$

PRED(l) is another evaluation criterion and shows the percentage of the estimates l percent different from the actual value. This parameter is defined in Equation7; in which N is the total number of reviewed studies and A the number of projects with *MRE* of less than l. The usual value for$l$ is 0.25, in this research too, *PRED(0.25)* was used. All of the criteria measure the accuracy of the estimation method; however, MMRE must be as small as and *PRED(0.25)* as big as, possible.

$$MMRE = \frac{\sum_{i=1}^{N} MRE}{N} \quad (6)$$

Table 4: Cross correlations analysis for ISBSG dataset

	InpCont	OutCont	EnqCont	FileCont	IntCont	AFP	NorEffort
InpCont	1	0.6036	0.6581	0.3875	0.2970	0.8779	0.7059*
OutCont		1	0.4298	0.4274	0.4219	0.7465	0.4865
EnqCont			1	0.2271	0.1962	0.7360	0.5555*
FileCont				1	0.3970	0.6055	0.3074
IntCont					1	0.4710	0.2927
AFP						1	0.6538*
NorEffort							1

Table 5: Cross correlations analysis for NASA dataset

	DL	M	Effort
DL	1	0.2715	0.9814*
M		1	0.2135
Effort			1

Table 6: Cross correlations analysis for Artificial datasets

	X1		X2		X3		Y	
	Moderate	Severe	Moderate	Severe	Moderate	Severe	Moderate	Severe
X1	1	1	-0.1269	0.0959	-0.0421	-0.1141	0.6877*	0.4048*
X2			1	1	-0.0213	-0.0422	-0.0330	-0.1758
X3					1	1	0.4961	0.1855
Y							1	1

$$PRED(l) = \frac{A}{N} \quad (7)$$

Correlation Analysis: Correlation analysis a statistical technique by which the dependency levels between different variables can be achieved. The equations' outputs of this analysis are in range of [-1,1]; a closer value to one indicates a stronger direct relationship between the two variables [38]. The advantages of this method include the speed and simplicity of implementation, interpretation and finding the relationships between the variables and one of its disadvantages is the inability to process categorical variables. Accordingly, in this paper, two basic considerations are made:

- Do we consider all the features of all datasets?
- Only features affecting the final effort value are selected. In the latter case, we are applying a type of feature selection that diminishes the size of the dataset.

At 0.05 level of significance, we performed spearman rank ordercross correlations analysis on the variables of each dataset to obtain the degree of correlation and dependency between independent variables and the effort [34]. The high values of this analysis indicate high dependency between variables and the values closer to 0 represent the relative independence of the considered variables. The analysis results using ISBSG, NASA and Artificial datasets are respectively presented in Tables 4, 5 and 6.

Each table presents all the potential correlations; however, we are only focused on the effort value. For better comparison, the highest values of each table (most effective features) are marked with star symbol. As we can see, for ISBSG dataset, the most important features (the highest correlation values) are respectively InpCont, AFP and EnqCont, which are used in the following comparisons. Since this dataset contains a variety of features, we only consider variables with correlation level higher than 0.5. Another interesting point is the considerable correlation of these three variables with one another; according to Table 4, it is clear that there is a significant relationship between the records of this dataset.

Moreover, according to Table 5, the most effective feature of NASA is *DL* with correlation value 0.9814, which indicates strong correlation with the effort value. In addition, the weak correlation between *DL* and *M* (0.2715) indicates their independence. Furthermore, the high value of *DL* shows that simpler estimation models (with less redundant features) can be made for this