MACHINE LEARNING TECHNIQUES IN SOFTWARE EFFORT ESTIMATION USING COCOMO DATASET

Sonam Bhatia[1], , Varinder Kaur Attri[2]
[1](Dept .of CSE, GNDU RC Jalandhar, India)
[2](Dept .of CSE, GNDU RC Jalandhar, India)

Abstract-

One of the most important tasks in software planning and management is estimation of the effort .Software has played an crucial model in software engineering and development for , complex systems. Reliable estimating the software size, cost, effort and schedule greatest challenge for software developers today. Overestimates and underestimates have direct impact for causing damage to software companies. In this paper we introduce a method based on machine learning technique. Linear regression and Multiperceptron are the most popular machine techniques for software development effort estimation. In this paper linear regression and multiperceptron have been used to predict the early stage effort estimations using the COCOMO dataset. It has been found that multiperceptron is able to predict the early stage efforts more efficiently in comparison to the linear regression models.

Keywords: Machine learning, Linear regression , feed forward neural network

I INTRODUCTION

Effective software evaluates the information needed to design a workable software development plan. How efficiently the project is evaluated is ultimately the key to the project's success. The size of the project is one of most indicators which should be noticed by developer. An effective software estimate provides important information for making project decisions, projecting outcomes, and defining goals [1]. Several indicators should be considered to estimate the software size and effort. The estimation of effort and cost depends on the reliable estimation of the size. The effort and cost estimations are biggest issue in the software projects. [2] Reliable software cost estimates are critical to both developers and customers. [3] There are great efforts and contributions to measure the size of a software system and evaluate the effort required to develop it. Size has direct affect on development effort and project management [4]. Many software effort estimation techniques have been proposed to evaluate their estimation computation. Many widely used approaches involve the estimation by expert analogy-based estimation [5] [6] [7].

Good evaluation plays essential role in the management of software projects. Many approaches proposed for effort estimation, including technique depends on machine learning, provides evaluate of the effort for a novel project [9] Machine learning in this new field, is illustrating the promise of developing accurate estimates. Machine learning system "tells" how to evaluate from training set of finished projects. [13] By applying learning approaches project managers and experts can take less time to evaluate software project effort and more time on more important challenge in leaving the project in time to customers.[14]

This paper is organized as follows. In Section 2 presents reason for difficulty for effort estimation .Section 3 highlights the literature survey the . In Section 4, describes the problem statement. Section 5 we present the method for effort estimation. The experiments and results are discussed in Section 6. Finally, Section 7 presents the conclusions.

II WHY SOFTWRAE EFFORT ESTIMATION DIFFICULT?

It is difficult to accurate estimate the effort due to the following reasons [12]:

1. Deficiency of historical database for size measurement
2. Factors that affect effort and productivity are not understood well. Their relation is also need to analyze.
3. Uneducated and untrained staff

III. LITERATURE REVIEW ON SOFTWARE EFFORT ESTIMATION TECHNIQUES

K.P.Manju in 2014[18] focused on better accurate estimation results based use case diagrams are given as input for software size and use case points is taken as output. A linear regression model with exponential transformation is used to find out there relationship between the variables that includes software size and effort for improving the accuracy in software effort estimation.

V. Anandhiin 2014 [19] explained the regression algorithms like M5 algorithm and Linear Regression in Software Effort Estimation using COCOMO dataset is evaluated. Simulation results demonstrate that the errors such as MMRE and MdMRE of M5 algorithm is less than linear regression in forecasting by 80.20 and 45.30 percentage respectively.

Jyoti Shivhare in 2014 [17] presented a paper and described an technique for estimation based upon various feature selection and machine learning techniques for non-quantitative data and is investigated in two phases. In the first phase of method three feature selection techniques, such as Rough-Reduct, RSA-Rank and Info Gain, are applied to the dataset to find the optimal feature set. The second phase include effort estimation for reduced dataset using machine learning techniques like FFNN, RBFN, FLANN, LMNN, NBC, CART and SVC .

Sumeet Kaur Sehra in 2011 [15] explained comparison; genetic programming can be used to fit complex functions and can be easily interpreted. Genetic Programming can find a more advanced mathematical function between KLOC and effort. Particle Swarm Optimization alone gives almost same results as basic models

Ch. Satyananda in 2009 [5]focused on COCOMO dataset and the experimental part of the study illustrates the approach and compares it with the standard version of the COCOMO. It has been found that Gaussian function is performing better than the trapezoidal function, as it demonstrates a smoother transition in its intervals, and the achieved results were closer to the actual effort.

Petrônio L. Braga in 2007[10] presented the strength and weakness of various software cost estimation methods. It also focuses on some of the relevant reasons that cause inaccurate estimation. In this paper a comparative analysis among existing popular models are performed and the performance is analyzed and compared in terms MMRE (Mean Magnitude of Relative Error) and PRED

Evandro N. Regoli [10] in 2003 explored two ML techniques, GP and NN. Author described that both techniques perform well in the regression problem. GP is able to investigate the correct functional equation that fits the data and its appropriate numerical coefficients. NN gives a net that express a complete mathematical formula, without a direct interpretation.

IV. PROBLEM STATEMENT

In order to fulfill the gaps machine learning techniques have been developed which led to estimate more accurately but as per there exist no perfect technique till now to estimate a size of software so it is one of the biggest problem need to be solved. Therefore, an accurate software effort estimation model is highly required in software project management.

V. MACHINE LEARNING TECHINQUES

The area of Machine Learning (ML) is used for evaluation methods that experiment various forms of learning, in particular approach capable of inducing knowledge from examples or data. [20]

A ESTIMATION TECHINQUES

1) LINEAR REGRESSION

Regressions techniques are used to predict software evaluate accuracy for evaluation and validation. A Regression Analysis is used to view the affect of independent variables on the dependent variable. The idea is to see that how much dependent variable is dependent upon the independent variables. Linear regression is a very grateful statistical technique [19]. Linear models can be used for prediction or to evaluate whether there is a linear interrelationship correlation between two numerical variables. A linear regression model with exponential transformation is used to predict out the relation between the variables that involve software size and effort for raising the reliability in software effort estimation. [18]We change the value of independent variable and see the resulting change in dependent variable. The objective is to find out at what extent dependent variable is described by using independent variable. In a Simple Linear Regression, we have one independent and one dependent variable . Mathematically, it can be written as:

Y = a + bX + C Where Y: Dependent variable X: Independent variable
b: Coefficient of variable X
a: Y intercept
C: Constant
In a Multiple Linear Regression Analysis, more than one independent variable is used to describe the change in dependent variable .
C: Constant

2) MULTI-PERCEPTRON
A neural net is a machine learning technique. Neural networks are presented in layers each involving of neurons or Process elements that is interconnected. The neurons or perceptrons compute a weighted sum of their inputs, generating an output. The first layer is act as the input layer; the last layer is act as the output layer and the layers between are hidden layers. Neural networks are nets of processing module that are able to learn the mapping existent between input and output data.[15][16].It is also act as as Feed Forward Neural Network (FFNN). A Feed Forward Neural Network (FFNN) works with back propagation learning algorithm is used to compute the effort. To develop FFNN, artificial neurons, also called nodes, are interconnected in the form of layers.MLF neural networks, with a back-propagation learning algorithm, are the famous neural networks.
Multi layer perceptron Can easily be predicted .They have small memory requirements and efficiently classify new data and exhibit good generalized capability. It is Computationally expensive learning approach because huge number of circulation needed for learning, that are not suitable for real-time learning.There is a scaling problem ,it is hard to scale.

B .VARIOUS CRITERIONS FOR SOFTWARE EFFORT

1) Correlation Coefficient*:* Correlation measures of the strength of a relationship between two variables. The larger the value of correlation coefficient, the stronger the relationship
2) MAE:It stands for Mean Absolute error.It measures of how far the estimates are from actual values.
3) RMSE:It stands Root Mean Square Error: It is frequently used measure of differences between values predicted by a model or estimator and the values actually observed from the thing being modelled or estimated
4) RAE: It stands for Relative absolute Error. Relative absolute Error (RAE) takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor
5) RRSE:It stands for Root Relative Squared Error[10]

VI. EXPERIMENTAL STUDY
In carrying out our experiments, we have COCOMO Dataset for software effort estimation is used to evaluate performance of the Multi layer perceptron and Linear Regression. The experimental setup consists 17 attributes and 81 instances.

Table I: Attribute of dataset

Attribute	**Description**
RELY	Required software reliability
DATA	Data base size
CPLX	Process complexity
TIME	Time constraint for CPU
STOR	Main memory constant

VIRT	Machine volatility
TURN	Turnaround time
ACAP	Analyst capability
AEXP	Application experience
PCAP	Programmers capability
VEXP	Virtual machine experience
LEXP	Language experience
MODP	Modern programming practice
TOOL	Use of software tools
SCED	Schedule constraint
LOC	Lines of code
ACT-EFFORT	Actual effort

Table II shows the results from the dataset.The model with the lower RMSE, RAE, RRSE, MAE and the higher correlation coefficient is considered to be the best model for estimation.

Table II: Comparison of Models

Evaluation criterions	**Linear Regression**	**Multi layer perceptron**
Correlation Coefficient	0.79994	0.8931
Mean Absolute error	247.0465	179.4526
Root Mean Square Error	431.768	310.3657
Relative absolute Error(%)	57.2976	41.6205
Root Relative Squared Error(%)	64.832	39.6376

The Figure1 shows the correlation cofficent ,mean absolute error,root mean square error,relative squared error achieved for COCOMO dataset using the linear regression and multi layer perceptron technique . The Results demonstrate that the errors such as RAE ,MAR,RRSE of Multilayer perceptron method is less than linear regression .

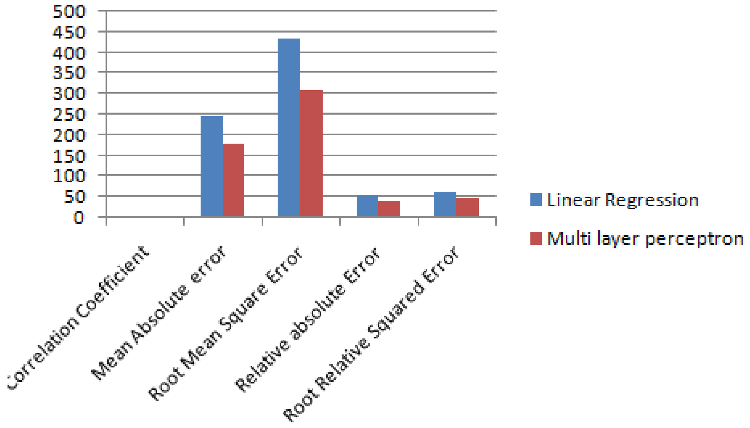

Figure 1:Result using linear regression and multilayer perceptron

The Figure2 shows the correlation cofficent of multilayer perceptron is more than linear regression

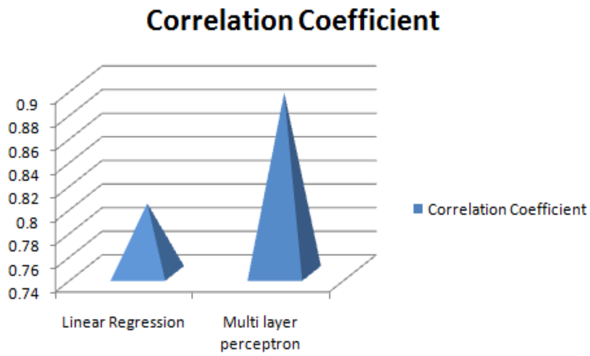

Figure 2:comparison of correlation cofficent

VII Conclusion

One of the important issues in software project management is accurate and reliable estimation of software effort especially in the early phase of software development. In this research we have made a comparative analysis of machine learning methods for predicting effort. Mean absolute error, root mean square error, relative squared error are used as evaluation criteria. The linear regression and multilayer perceptron is used for evaluation .These techniques have been compared in terms of accuracy. Research demonstrates that multilayer perceptron model is better than linear regression model.

.

REFRENCES

[1] Venus Marza and Mir Ali Seyyedi, "Fuzzy Multiple Regression Model for Estimating Software Development Time" *International Journal of Engineering Business Management, Vol. 1, No. 2 (2009), pp. 31-34*

[2] 1Vahid Khatibi, 2Dayang N. A. Jawawi, " Software Cost Estimation Methods: A Review" Journal of Emerging Trends in Computing and Information Sciences, *, Vol. 2, No. 1 (2011).*

[3] Jyoti G. Borade, " Software Project Effort and Cost Estimation Techniques International Journal of Advanced Research in Computer Science and Software Engineering" Volume 3, Issue 8, August 2013

[4] Luís M. Alves, "An Empirical Study on the Estimation of Software Development Effort with Use Case Points", IEEE 2013

[5] Ch. Satyananda Reddy, KVSVN Raju , "An Improved Fuzzy Approach for COCOMO's Effort Estimation using Gaussian Membership Function" JOURNAL OF SOFTWARE, VOL. 4, NO. 5, JULY 2009

[6] Jorgenson M, Sjoberg D.I.K., The impact of customer expectation on software development effort estimates.*International Journal of Project Management*, 22(4) :317–325.

[7] Chiu NH, Huang SJ, "The adjusted analogy-based software effort estimation based on similarity distances,"*Journal of Systems and Software*, Volume 80, Issue 4, April 2007, Pages 628-640.

[8] Topi Haapio , "*Improving Effort Management in Software Development Projects" ,2012*

[9] Petrônio L. Braga, "Software Effort Estimation using Machine Learning Techniques with Robust Confidence Intervals", 19th IEEE International Conference on Tools with Artificial Intelligence,2007

[10] Sweta Kumari , Shashank Pushkar, " Performance Analysis of the Software Cost Estimation Methods: A Review" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013

[11] Sohaib Shahid Bajwa, " Investigating the Nature of Relationship between Software Size and Development Effort"2009

[12] E. J. Pedhazur, 1997. Multiple Regression in Behavioral Research: Explanation and Prediction, Third Edition, Harcourt Brace College Publishers, ISBN: 0-03-072831-2.

[13] Yogesh Singh, Pradeep Kumar, "A REVIEW OF STUDIES ON MACHINE LEARNING TECHNIQUES" International Journal of Computer Science and Security, Volume (1) : Issue (1)

[14] B.V. Ajay Prakash1An Evaluation of Neural Networks Approaches used for Software Effort Estimation" *Association of Computer Electronics and Electrical Engineers, 2013*

[15] Sumeet Kaur Sehra, "SOFT COMPUTING TECHNIQUES FOR SOFTWARE PROJECT EFFORT ESTIMATION" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624.Vol 2, Issue 3, 2011, pp 160-167

[16] Daniel Svozil, "Introduction to multi-layer feed-forward neural networks"

[17] Jyoti Shivhare , "Effectiveness of Feature Selection and Machine Learning Techniques for Software EffortEstimation"2014

[18] K.P.Manju, "A Linear Regression Model with ExponentialTransformation for Software Effort Estimation" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering ,Vol. 3, Special Issue 2, April 2014

[19] V. Anandhi1 , R. Manicka Chezian, "REGRESSION TECHNIQUES IN SOFTWARE EFFORT ESTIMATION USING COCOMO DATASET" International Conference on Intelligent Computing Applications,2014

[20] Evandro N. Regolin Gustavo A. de Souza,"Exploring Machine Learning Techniques for Software Size Estimation" ,International Conference of the Chilean Computer Science Society (SCCC'03)1522-4902/03 $ 17.00 © 2003 IEEE