Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Predicting the Software Effort Estimation with Machine Learning Techniques and their comparision

Sonam Bhatia, Varinder Kaur Attri
Dept. of CSE, GNDU, Punjab,
India

Abstract— Software effort estimation is the one of the most enduring problem in software engineering. The role of software engineering cannot be ignored in development of the software. It is important for an industry to understand how to create error free software at fewer prices with less time and compute risk that software presents in daily life. Software has played a crucial model in software engineering and development for, complex systems. The dissertation aims empirically evaluates and compares the potentials of Machine learning approaches that have been used to predict the early stage effort estimations using the datasets. This paper focuses on Linear Regression, Multi layer perceptron and decision tree to evaluate the effort on two different data. .The performance of techniques is estimated based on evaluation criterions.In estimation techniques; it is very hard to determine which technique provides better estimation on which dataset.

Keywords— Effort , linear regression, Multi-Layer Perceptron ,Decision tree M5

I. INTRODUCTION

Effort estimation is one of the most essential activities in the development of the software .Estimation of software is difficult task in project planning and management process. [1]Software effort estimation is the one of the most enduring problem in software engineering. Software is the most expensive component in many computer based systems .A large amount of bugs creates huge difference between gain and loss during the estimation of effort. [2] Software effort estimation is the process of evaluating the most practical use of effort needed to produce or maintain software based on inadequate, unreliable input. Effort is usually evaluated in Person-Month. [3] Effort estimates acts as input for project plans. With the help of Software effort estimation it is easy to find resources which are used to complete the projects on time. Software effort estimation plays imperative role in the finalization of any project. Accurate computations normally lead towards the completion of the project on the right time. To manage the resources for developing the software, reliable estimation is very necessary. [4]

The world's economy affects a lot with software industry. There is a lot of pressure on industry to supply good quality software within time and budget control due to competitive market. So software industries required to make a balance between quality and effort. Unreliable estimation of effort deals with a great loss to the software industry. [5] Good estimates play a vital role in the management of software projects. Various methods are available for effort estimation of software project. The effort is the most significant factor that affecting the budget of a software project Overestimates and underestimates have direct impact for causing damage to software companies. It is the responsibility of the project manager to make accurate estimates of effort and cost. Estimating the effort with a high degree of efficiency is a matter which has not yet been solved and even the project manager has to deal with it since the beginning. There are different parameters can affect the effort estimation are size ,maintainence ,category[6][7]

Most techniques about software effort estimation use statistical approaches, which are not able to present reason and strong results. Machine learning methods play significance role in this filed because they can raise the efficiency of estimation by applying training rules of estimation and repeating the run cycles [8].

This paper is organized as follows: In section II we present work related to techniques. In section III we present research background. In section IV, research methodology is presented .In section V ,resukt and discussion is laid out. Conclusion is presented in section IV. In section V we discuss future scope.

II. RELATED WORK

The estimation techniques needed to develop software has been surveyed in this section. The work done by various authors has been compiled in this section.

performed accurately.

2010,Saleem Basha and Dhavachelvan P [6] described that SEER and machine learning techniques were reliable good at predicting the effort. But however they are not accurate because all the model lies in the term prediction, prediction never comes true is proved in this estimation models. Finally this paper concludes that the no model is best for all situations and environment [12]

2011,Olga Fedotova and Leonor Teixeria,Helena Alvelos [11] explained Software development organization that is used for the capability maturity model integrated .They described .software development organization evaluates the

effort used by the software based on the field of expert. They used stepwise linear regression technique and observed that linear regression perform better as compared to expert judgments. Value of mean magnitude relative error deviation and percentage relative error given by expert are 0,161and 74 % Value of mean magnitude relative error deviation and percentage relative error given by linear regression are 0,158and 79% .They also explained that testing team gave better outcomes as compared to expert judgment.

2011,Abdulbasit S. Banga [12] described the machine learning methods or algorithmic cost models .Author explained the benefits, limitations of every approach , models and the underlying manner in assembling cost estimates. Analyzing or comparison of many approaches of estimation or model are explained by author.

2011,Ruchika Malhotra and Ankita Jain [9] presented a paper and estimate ,compares the Linear Regression, ANN, Decision Tree, SVM , Bagging techniques on project of software dataset. The dataset which is taken from 499 projects are used .This dataset contain 19 features that we have to shortened ten features with the help of CFS method. The outcomes show that decision tree approach have Mean Magnitude Relative errors of 17% as compared to other approach Thus, the estimation or outcomes of decision tree approach is good rather than any other methods.

2012,Roheet Bhatnagar[17] used neural network approach and FIS approach to evaluate the effort .Author explained that linear regression neural network has less Mean magnitude relative error when compared with other neural networks FFBPNN Model gave value for MMRE is 12.96 LRNN Model gave value for MMRE was 11.45 and value for Mamdani FIS was 3.89 .When comparison of both linear regressions and fuzzy logic occurred then it is viewed that fuzzy logic gave better performance as compared to linear regression neural network for effort estimation.

2012,Ali Bou Nassif and Dr. Luiz Fernando Capretz [7] analyzed a model which involves linear regression, non linear regression, feed forward neural network, radial neural network and general regression neural network, tree boost model. Based on size of testing, training of data points they performed many observations to train or test the model. Author performed neural network against regression methods .Also performed neural network against other two models which compute the estimation from use case diagrams. It was observed that non linear regression gave better performance as compared to linear. Author also observed that the interrelationship between effort and size is non-linear. The GRNN model gave better result as compared to RBFNN.

2013, Mohd. Sadiq [10] explained the linear regression model for evaluating the software effort. Author described before estimating the effort of software project ,it is necessary to gain information about count of function point Also conducted that cost required to complete the function point was 0.1804 man –day The value for mean magnitude relative error is observed to be In study the value of the MMRE is found to be 0.1356.

2014,Jyoti Shivhare [13] performed estimation by several neural network and classification methods .Author explained methods for evaluation that are based many feature selection ,approaches of machine learning for data which is non quantitative .Author considered it in in two stages or steps . In the first step of method there are three feature selection approaches. These are Rough- Reduct, Rough set analysis and Information Gain are enforced to the dataset. These are used to determine the optimal feature set. In second stage machine learning approaches such as feed forward, radial bias function ,functional link neural network, LMNN, naïve bayes classifier ,CART, support vector machine are used for evaluation for reduced dataset.ANN machine learning techniques for USP05-FTdata.In order to achieved optimal pair ,these approaches are compared. The result shows that feed forward neural network neural network and naïve bias classifier gave better result as compared to other classification techniques.

2014, NehaSaini and Bushra Khalid [15] computed machine learning approaches for effort evaluation of the software. These techniques are decision trees, MLP, decision tables, perceptron , bagging, radial bias networks. These techniques are applied on heiatheiat dataset and miyazaki94 dataset . Authors explained that Decision trees approach is better for computing the effort. They explained that decision tree perform best instead of other models in term of MMRE value. The MMRE value for miyazak 94 dataset is 0.0436, Mean relative error value for decision tree is -0.036. For heiatheiat dataset MMRE value is 0.1074, Mean relative error is -0.00019.

2014, Berna Seref and Necaattin Barisci [16] performed effort estimation with the help of Multi layer perceptron and adaptive neuro fuzzy interference system. They analyzed the NASA dataset with 93 projects, Desharnais dataset with 77 projects .Mean magnitude relative error, Percentage relative error for NASA dataset are 0.18,0.8 when adaptive neuro fuzzy is being applied but the values for Desharnais were 0.94,0.18.It has been observed that ANFIS produce better outcomes as compare to multi layer perceptron. Mean magnitude relative error, Percentage relative error for Desharnais dataset are 0.94,0.18 when is being applied multi layer perceptron but the values for Desharnais were 1.23,0.09.It has been observed that ANFIS produce better outcomes as compare to multi layer perceptron. It showed that estimation for NASA Projects are near to actual efforts when Compared estimations for Desharnais Project.

III. RESEARCH BAKGROUND

A. Tool Used

There are many tools available for software effort estimation using machine learning techniques. Orange, Weka, MATLAB are available for effort prediction. Here we will use Weka tool for experiment of machine learning method for evaluation of performance measures for effort. Weka is used to resolve huge amount of issues such as classification, clustering, and neural networks.

WEKA stands for Waikato Environment for Knowledge .It is machine learning toolkit developed at Waikato University in New Zealand. It implements machine learning algorithm that are written in java. It is a open source software. It is used in research area, education, and software projects. The University of Waikto initiated the development of version of Waikato Environment for Knowledge in 1993 .

The decision was made to redevelop Waikto Environment for Knowledge from scratch in java involving exertion of modeling decision.[81] There are many version of Waikato Environment for Knowledge for example Weka 3.0.Weka 3.2.Weka 3.2 version provides graphical user interfaces. Tool based on Graphical user mainly utilized for preprocessing, computation method and provide platform for comparison of various machine learning methods. The techniques are applied directly on data sets. Waikto Environment for Knowledge tool implement the algorithm for classification, clustering. Methods or approaches of Waikto Environment for Knowledge are based on the fact that data is present as a relation and also based on that every data point is associated by fixed number of attributes. [78]

B. Data Collection

We used dataset to evaluate the reliability of software with help of machine learning methods. These datasets are present in promise data repository. These dataset are:

COCOMO Dataset: It is a is a PROMISE Repository dataset in order to encourage repeatable, refutable or improvable predictive model for Software Engineering .It consists of 60 NASA projects taken from different centre's. The experimental setup consists17 attributes and 81 instances. Rows of dataset represent the project, column represent the attributes.

Desharnais dataset: This dataset consist 81 instances and 12 attributes. This data will be stored in table form. CSV file format is used to store data. CSV file is created using MS EXCEL and to store data. After storage of data this excel file is saved with .csv extension to use it for data mining process.

C. *Evaluation Measures*

There are various evaluation measures for accuracy of the development effort.

- **Correlation Coefficient**
  Correlation measures of the strength of a relationship between two variables. The larger the value of correlation coefficient, the stronger the relationship is.
- **MAE**
  It stands for Mean Absolute error. It measures of how far the estimates are from actual values.
- **RMSE**: It stands Root Mean Square Error: It is used for magnitude of deviation between values presumed by estimator and the values actually examined from the thing being modeled .
- **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
- **RRSE**:It stands for Root Relative Squared Error

IV. RESERCH METHODOLOGY

In this paper we are using machine learning approach like linear regression , multilayer perceptron, decision tree.in order to predict effort. WEKA tool is used for performance measures and CSV file is used for datasets. Linear regression ,Multiperceptron,and decision tree was used to perform evaluation for both the datasets.

A. Linear Regression

Linear regression is the highest extensively used of all statistical approaches. It is the linear interrelation between variables. This is an way for modeling the interrelation between a scalar dependent variable which is denoted as y and independent variable that is denoted with X. [9] [10]. When there is more than one independent variable, the process is called multiple linear regressions [11] Linear model considers the relationships between variables are straight-line relationships.

B. Multi layer perceptron

A multilayer perceptron (MLP) is a feed forward artificial neural network for structuring the neurons is known as multi-layer neural network. Artificial neurons are interconnected in the form of layers. It involves multiple layers of perceptron with each layer fully attached to other layers of network and the network view as mesh network. layers of network and the network view as mesh network. The first layer called the input layer neurons that express the set of input variables. The output layer express the output variable which is the actual effort required to end the project. The connections between the neurons have weighted numerical inputs associated with them. The layers between two layers are known as hidden layers .All the neurons of one layer generate some output, which acts as input to the next layer. This `next layer' can be either the hidden layer or the output layer [12] [13]. It uses back propagation learning algorithm is used to compute the effort [13]

C. Decision tree

Decision tree algorithms are mostly used for classification task. M5P is powerful because it implements as much decision trees as linear regression for predicting a continuous variable. This algorithm is a multivariate tree algorithm which is appropriate for noise removal and also applies for large database The M5P Introduced by Quinlan, the model tree technique (M5) can be recognized as an extension to CART. A model tree will fit a linear regression to the observations at each leaf rather of allowing a single value like CART. Mistakes reduction can be done by using this algorithm[14].

From data .csv file was created and it is loaded into WEKA explorer. WEKA explorer contains preprocess panel, by clicking on this you can see the open file tab; by clicking on it you will load the file in WEKA. Now click on the classify panel then click on techniques .There are number of techniques choose Linear regression ,Multiperceptron and decision

tree one by one to perform evaluation. Now under test option 10 folds cross validation is chosen as our evaluation method. Now we will perform evaluation by executing these machine learning approaches.

V. RESULT AND DISCUSSION

A .Comparison of techniques

Implementation is done using WEKA tool. To perform evaluation 10 folds cross validation method is used. Comparison of Linear regression , Multiperceptron, and decision tree is performed using 10 folds cross validation on both the datasets. Results obtained from this evaluation are given below.

Performance comparison: Performance comparisons of techniques of COCOMO dataset 10 folds cross validation is given in Table I. From this comparison it is clear From TABLE. I ,it is observed that decision tree i.e M5 give better result as compared to other techniques.

Table I Performance and Comparisons of Techniques for COCOMO dataset

Evaluation criterions	Linear Regression	Multi layer perceptron	Decision tree
Correlation Coefficient	0.79994	0.8931	0.9152
Mean Absolute error	247.0465	179.4526	157.1147
Root Mean Square Error	431.768	310.3657	263.9787
Relative absolute Error	57.2976	41.6205	36.4397
Root Relative Squared Error	64.832	46.6029	39.6376

The Value of mean absolute error, root mean absolute error, root mean square error, root relative square error are observed to be 157.11,263.97,36.43,39.6376%.On the basis of these values it is found that decision tree give best outcomes and to be effective in evaluating effort

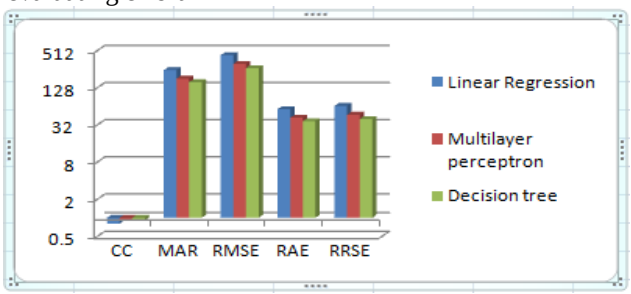

Figure 1: Graph representing comparision of techniques for COCOMO dataset

The Fig 1 shows the performance measures for COCOMO dataset with machine learning methods. From figure we viewed that decision tree provides better result.

Another Observations or experiment involves the Desharnais dataset for computation the performance of machine learning approaches. Performance measures of machine learning with 10 folds cross validation is given in TABLE II. From comparison it is clear that linear regression method provides good results because it has high correlation coefficients and low all other performance measures. The value of correlation coefficients and MAR, RMSE, RASE.RAE, RRSE for linear regression are0.7277, 2134.5209, 3033.6563, 66.5719 %, 67.4574 %

Table II. Performance and Comparison of Techniques for Desharnais Dataset

Evaluation criterions	Linear Regression	Multi layer perceptron	Decision tree
Correlation Coefficient	0.7277	0.6252	0.6955
Mean Absolute error	2134.5209	3075.1232	2204.4842
Root Mean Square Error	3033.6563	4616.8678	3192.9174
Relative absolute Error	66.5719	95.9077	68.754
Root Relative Squared Error	67.4574	102.662	70.998

The Figure 2. shows the performance measures for Desharnais dataset with machine learning methods. From figure we viewed that linear regression provides better result.

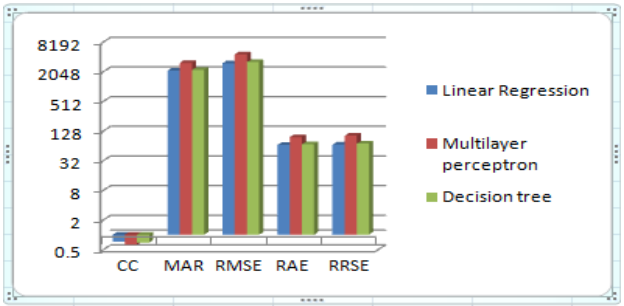

Figure 2.Graph for comparison of performance measures for Desharnais dataset

B. Comparison of Datasets

Weka is learning tool that provide large spectrum of machine learning techniques. In our implementation weka is used for comparison between the performance measures of different classifiers like linear regression, multi layer perceptron and M5 decision tree for both the datasets. The comparison is used to find out the best technique for effort estimation from both the dataset.From the TABLE III, performance measures shows that decision tree give better results for the COCOMO dataset when compared to Desharnais

Table III Comparisons of Techniques for Desharnais dataset , COCOMO dataset

	COCOMO DATASET			DESHARNAIS DATASET		
Performance Measures	**Linear Regression**	**Multi layer perceptron**	**Decision tree**	**Linear Regression**	**Multi layer perceptron**	**Decision tree**
Correlation Coefficient	0.79994	0.8931	0.9152	0.7277	0.6252	0.6955
Mean Absolute error	247.0465	179.4526	157.1147	2134.5209	3075.1232	2204.4842
Root Mean Square Error	431.768	310.3657	263.9787	3033.6563	4616.8678	3192.9174
Relative absolute Error	57.2976	41.6205	36.4397	66.5719	95.9077	68.754
Root Relative Squared Error	64.832	46.6029	39.6376	67.4574	102.662	70.9987

VI. CONCLUSION

A software supplier organization strives to evaluate the effort needed in constructing software as accurately as accessible to set out the project budget and plans and the achievement of resource appropriation. Software effort estimation is a crucial part of a successful software development**.** As software becomes more complex and its field dramatically rise, the significance of research on developing approaches for appraisal software development time has always raised, so accurate estimation or appraisal is the main goal of software managers for decreasing risks of projects

.The outcomes obtained from applying linear regression ,multi layer perceptron ,decision tree to COCOMO dataset ,showed that decision tree M5 computed better estimation results than other approaches using performance measures MAR =157,RAE=36 %,RRSE=39%. When these techniques utilized on second dataset i.e Desharnais, results showed that linear regression performed better estimation than multi layer ,decision tree using performance measures MAR=2134.5209,RAE=66%,RRSE=67%.

In the last we have made a comparative analysis of both the datasets on same techniques and find the better one to be implemented. The results showed that decision tree of COCOMO evaluated better outcomes or results.

VII. FUTURE WORK

Many applications will be designed which have expected to have less size or effort so that software complexity can be reduced. The future work is to study new software effort estimation methods and models that can be help us to easily understand the effort estimation process of the software. The work can be further done by selecting a combination of machine learning technique which provides better results.

.ACKNOWLEDGMENT

First of all I express my sincerest debt of gratitude to the Almighty God who always supports me in my endeavors.

I would like to thank teachers for their encouragement and support. Then, I would like to thank my family and my friends. I am thankful to all those who helped me in one way or the other at every stage of my work.

REFERENCES

[1] Jonathan Lee,Wen-Tin Lee,Jong-Yih Kuo, "Fuzzy Logic as a Basic for Use Case Point Estimation" IEEE International Conference on Fuzzy Systems ,June 27-30, Taipei, Taiwan, 2011

[2] Khaled Hamdan, Mohamed Madi," Software Project Effort: Different Methods of Estimation", International Conference on Communications and Information Technology (ICCIT), Aqaba.,IEEE,2011

[3] Alhad Vinayak Sapre, " Feasibility of Automated Estimation of Software Development Effort in Agile Environments",2011

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

[5] Mamoona Humayun and Cui Gang, "Estimating Effort in Global Software DevelopmentProjects Using Machine Learning Techniques" International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012

[6] Saleem Basha,Dhavachelvan "Analysis of Empirical Software Effort Estimation Models" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010

[7] Ali Bou Nassif, Dr. Luiz Fernando Capretz "Software Size and Effort Estimation from Use Case Diagrams Using Regression and Soft Computing Models"May2012

[8] Vahid Khatibi, Dayang N. A. Jawawi , " Software Cost Estimation Methods: A Review" Journal of Emerging Trends in Computing and Information Sciences,2011

[9] Ruchika Malhotra, "Software Effort Prediction using Statistical and Machine Learning Methods", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.1, January 2011

[10] Mohd. Sadiq, Aleem Ali, Syed Uvaid Ullah, Shadab Khan, and Qamar Alam, " Prediction of Software Project Effort Using Linear Regression Model"International Journal of Information and Electronics Engineering, Vol. 3,No. 3, May 2013

[11] Olga Fedotova,Leonor Teixeria,Helena Alvelos , "Software Effort Estimation with Multiple Linear Regression: review and practical application Journal Of Information Science And Engineerin,2011

[12] Abdulbasit S. Banga "Software Estimation Techniques" National Conference; INDIACom-2011Computing for Nation Development, March 10 – 11, 2011

[13] Jyoti Shivhare, "Effectiveness of Feature Selection and Machine Learning Techniques for Software Effort Estimation", June 2014

[14] Aditya Polumetla , " Machine Learning Methods For The Detection of RWIS Sensor Malfunctions",2008

[15] NehaSaini1 ,Bushra Khalid, " Empirical Evaluation of machine learning techniques for software effort estimation" Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661

[16] Berna Seref and Necaattin Barisci , "Software Effort Estimation Using Multilayer Perceptron and Adaptive Neuro Fuzzy Inference System" International Journal of Innovation, Management and Technology, Vol. 5, No. 5, October 2014

[17] Roheet Bhatnagar1 and Mrinal Kanti Ghose ,"Comparing Soft Computing Techniques For Early Stage Software Development Effort Estimation" International Journal of Software Engineering & Applications (IJSEA), Vol.3, No.2, March 2012

