ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Implementing Decision Tree for Software Development Effort Estimation of Software Project

Sonam Bhatia, Varinder Kaur Attri

Student, Dept. of CSE, GNDU, RC, Jalandhar, India

Assistant Professor, Dept. of CSE, GNDU, RC, Jalandhar, India

ABSTRACT: Effort estimation is one of the biggest problems faced by software industry. In software planning estimation of the effort is one of the most critical responsibilities. It is necessary to have good effort estimation in order to conduct well budget. The accuracy of the effort estimation of software projects is vital for the competitiveness of software companies. For the forecasting of software effort, it is important to select the correct software effort estimation techniques. Inaccurate effort estimation can be risky to an IT industry's economics and certainty due to poor quality or trait and stakeholder's disapproval with the software product. This paper presents M5P decision tree Technique, for effort evaluation in the field of software development**.**

KEYWORDS: Effort estimation, Decision tree, M5P , Machine learning

I .INTRODUCTION

Software effort estimation is the forecasting about the amount of effort needed to make a software system and its duration [1] Good estimates play a very important role in the management of software projects.[2] . The effort is the most important cause that affecting the budget of a project. Estimating the effort with a high degree of accuracy is a issue which has not yet been solved and even the project manager has to deal with it since the beginning. Several parameters can affect the effort estimation. These parameters Incorporate Size, Category, Personnel Attributes, Complexity [3] Most of the effort estimation metrics takes the input as the software size, which can be measured with function point, LOC, object point. A number of models have been enlarged to provide the relation between size and effort. [16]
SLOC is typically used to predict the amount of effort that will be needed to establish a program, as well as to valuation programming productivity or maintainability once the software is developed[11] Effort is measured in terms of person months and duration.[4]. More recently attention has turned to a variety of machine learning techniques to predict software development effort [7][8]. Most of the projects are break down due to imprecise estimated effort, so the success of any software project depends on an initial and accurate effort estimation.[9] the purpose of Machine Learning is to provide increasing levels of automation in the knowledge engineering process, replacing much time consuming human activity with automatic techniques that improve reliability or efficiency by observing and manipulating regularities in training data.[5]

There are many reason for vary of effort estimation. These are Project approval, project management, defining of project task etc. The field of Machine Learning (ML) is devoted to develop computational methods that implement various forms of learning, in particular mechanisms capable of inducing knowledge from examples or data [10]. An important requirement is that the learning system should be able to deal with imperfections of the data. Many methods have been explored for software effort estimation, consisting traditional methods such as the COCOMO and, more recently, machine learning techniques such Linear regression, Multi-Layer Perceptron, Decision tree[2] Machine learning methods have been exploited to generate better software products ,to be part of software product and to make software development process more convenient and adequate .[6]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

The remainder of this paper is laid out as follows. Section II describes importance for effort estimation and performance measures. Section III,describes related work. In Section IV, explains how decision tree can be used and implementation of M5P algorithm over sample data. Section V gives the result. Final section gives conclusion on this survey

II. IMPORTANCE FOR EFFORT ESTIMATION AND PERFORMANCE MEASURES

A .Importance

Effort estimation is necessary for many people and different departments in an organization. At various point of project lifecycle well-defined effort estimation is essential. The computation of the effort might be used as input to project plans, determining the budget and other important procedure needed for the successful release of the software. The progress or failure of projects depends on the authenticity or reliability of effort and schedule evaluations, among other things. Early effort estimation also assists the project manager to investigate whether the available resource is effective to complete the project. As software applications have grown in size and significance, the need for reliability in software cost estimating has grown, too.

B. Performance measures

Correlation measures of the strength of a relationship between two variables, Mean Absolute error measures of how far the estimates are from actual values, Relative absolute Error (RAE) takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor. Root Mean Square Error (RMSE): RMSE evaluates the difference between value estimated by a model and the value actually observed. [9]

III. RELATED WORK

Jyoti Shivhare[9] in 2014 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[7] in 2011 described that the Radial basis neural network gives more reliable results as compared to intermediate COCOMO Model and fuzzifying size and cost drivers by using Gaussian MF. The accuracy of effort estimation can be improved and the estimated effort is very close to the actual effort.Also explained genetic programming based effort model provides results which are more robust and accurate.

NehaSaini[14] in 2014 evaluated various machine learning techniques for software effort estimation like bagging, decision trees, decision tables, multilayer perceptron and RBF networks. Two different datasets i.e. heiatheiat dataset and miyazaki94 dataset have been used in research. Decision trees are good for evaluating the software effort. Also author described that Decision trees perform best among a other models in term of MMRE value.

Karel Dejaeger[13] in 2012 presented a paper and explained that ordinary least squares regression in combination with a logarithmic transformation performs best. By selecting a subset of highly predictive attributes, typically a significant increase in estimation accuracy can be achieved. These results also demonstrate that data mining approches can make a valuable comission to the set of software effort estimation techniques, but should not change expert judgment.

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.

Ruchika Malhotra[4] in 2011 presented a paper and estiamte ,compares the potential of Linear Regression, Artificial Neural Network, Decision Tree, Support Vector Machine and Bagging on software project dataset. The dataset is obtained from 499 projects. The results show that Mean Magnitude Relative error of decision tree method is only 17.06%. Thus, the performance of decision tree method is better than all the other compared methods.

Sweta Kumari[16]in 2013 provided a comparative study on support vector regression (SVR), Intermediate COCOMO and Multiple Objective Particle Swarm Optimization (MOPSO) model for effort estimation and SVR gives better results.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

IV PROPOSED MODEL

Proposed model includes four steps .The steps are indentifying the problem domain, scanning data, partition data into test and training or classification.

A .Data preprocessing

There are 17 attributes used in our data set and a brief description about each is presented in table1.

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

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering, and association rules.

B .Decision Tree

A decision tree is a logical model that is contributes in operations research, specifically in decision analysis [9]. Decision tree is a kind of tool to come out with a decision on the basis of some conditions and their possible consequences.[16] Decision tree is a procedure used for classification and regression [15].Decision tree is a flowchart like tree structure, where each internal node stand for a test on an attribute, each branch express an outcome of the test, and each leaf node holds a class label. The root node is the topmost node in a tree [15]

Decision trees are generated from training data in a top down, general to specific direction. The initial state of tree is root node that is assigned all examples from training the training set. If it is case that all the examples belong to same class then no further decision need to be made to partition the examples and the solution is complete. If example at this node belongs to two or more classes then test is made at node that will result in split. The process is recursively repeated for each intermediate node until completely discriminating tree is obtained.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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

fit a linear regression to the observations at each leaf rather of allowing a single value like CART. The M5P algorithm has three stages: building a tree, pruning the tree and smoothing. [4]

V.RESULT

A .M5P Implimentation result

Using weka tool the classification algorithm is used to perform the experiments on COCOMO dataset .From the given dataset by using M5P algorithm, the following rules were produced on the basis of data

```
=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

LOC <= 100.5 :
|   LOC <= 32.55 : LM1 (32/4.183%)
|   LOC >  32.55 : LM2 (15/7.505%)
LOC >  100.5 : LM3 (13/49.382%)

LM num: 1
ACT_EFFORT =
        80.816 * TIME=Very_High,High
        + 37.0466 * VIRT=Low,High
        + 97.0766 * VEXP=Nominal,High
        + 24.7874 * TOOL=High,Very_High,Low,Very_Low
        + 3.9701 * LOC
        - 137.5827


LM num: 2
ACT_EFFORT =
        80.816 * TIME=Very_High,High
        + 37.0466 * VIRT=Low,High
        + 162.7023 * VEXP=Nominal,High
        + 82.9084 * TOOL=High,Very_High,Low,Very_Low
        + 4.0844 * LOC
        - 161.764

LM num: 3
ACT_EFFORT =
        520.4254 * TIME=Very_High,High
        + 365.9082 * VIRT=Low,High
        + 168.2373 * VEXP=Nominal,High
        + 7.1874 * LOC
        - 961.138

Number of Rules : 3

Time taken to build model: 0.17 seconds
```

Fig 1: RULES GENERATED BY TREE

B. Performance Measure

By using algorithm correlation coefficients, mean absolute error, root mean squared error, relative absolute error, and root relative absolute squared error were measured

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.922
Mean absolute error                    150.9841
Root mean squared error                252.8864
Relative absolute error                 35.0178 %
Root relative squared error             37.9721 %
Total Number of Instances               60
```

Fig2: Performance measures

On the basis of rules the following tree was generated for our data set .

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

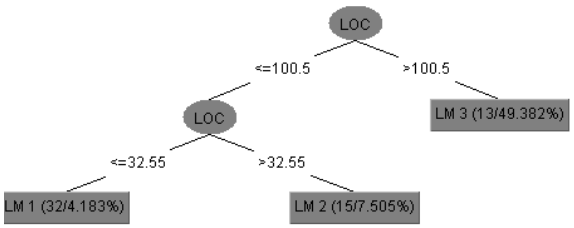

Fig 3: Tree generated

VI .DISCUSSION OF RESULTS

In this analysis,we used M5P Classification tree on 60 instances of COCOMO dataset .Correlation measures was 0.922, Mean Absolute error measures was 150.9841, Relative absolute Error (RAE) was 35.0178 %and. Root Mean Square Error (RMSE) was 252.8864 ,root relative square error was 37.9721%

VII. CONCLUSION

Effort estimation is greatest problem faced by software industry. Effort is very closely related to size of the software and cost of the software hence it is very important for software industry to reduce the effort .Data mining technique like decision tree can be used to study the effort for software. The result of our paper focuses on the implementation of M5Pdecision tree. It is assumed that with better characteristic of decision tree can generate a specialized method to monitor the effort or performance measures and hence take necessary steps to reduce the effort.

ACKNOWLEDGEMENT

We would like to thank acknowledge almighty for his constant blessings. Then we like to thank our family and friends for helping and supporting us throughout the making of this paper

REFERENCES

[1] Olga , “Software Effort Estimation with Multiple Linear Regression: review and practical application Journal Of Information Science And Engineerin(2011)

[2] Petrônio L. Braga and Adriano L. I. Oliveira, “,Software Effort Estimation using Machine Learning Techniques with Robust Confidence Intervals”,19th IEEE International Conference on Tools with Artificial Intelligence

[3] Ali Bou Nassif, “ Software Size and Effort Estimation from Use Case Diagrams Using Regression and Soft Computing Models” ,MAY 2012

[4] 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

[5] Yogesh Singh, Pradeep Kumar Bhatia & Omprakash Sangwan, “A Review Of Studies On Machine Learning Techniques” International Journal of Computer Science and Security, Volume (1) : Issue (1)

[6] Zhang, “Advances in Machine Learning Applications in Software Engineering”

[7] Sumeet Kaur Sehra1, Yadwinder Singh Brar2, and Navdeep Kaur3 , “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

[8] Prasad Reddy P.V.G.D, Sudha K.R, Rama Sree P and Ramesh “Software Effort Estimation using Radial Basis and Generalized Regression Neural Networks”, Journal of Computing, Volume 2, Issue 5, pp 87-92,. 2010

[9] Jyoti Shivhare, “Effectiveness of Feature Selection and Machine Learning Techniques for Software Effort Estimation” June 2014

[10] 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

[11] Kaushal Bhatt, Vinit Tarey, Pushpraj Patel , “Analysis Of Source Lines Of Code(SLOC) Metric”, International Journal of Emerging Technology and Advanced Engineering(ISSN 2250-2459, Volume 2, Issue 5, May 2012)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

[12] Geetika Batra, Kuntal Barua , "A Review on Cost and Effort Estimation Approach for Software Development" International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 4, October 2013
[13] NehaSaini ,Bushra Khalid, "Empirical Evaluation of machine learning techniques for software effort estimation" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. IX (Feb. 2014), PP 34-3
[14] Jiawei Han Data Mining:Concepts and Techniques Second Edition,2011
[15] Sweta Kumari "Comparison and Analysis of Different Software Cost Estimation Methods"(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.1, 2013
[16] http://www.qualitymanagementconference.com/effort_estimation.php

BIOGRAPHY

Sonam Bhatia received B.Tech degree in Information Technology from Punjab Technical University,Jalandhar,India,in 2013,pursuing M.tech in Computer Science Engineering from Guru Nanak Dev University,Amritsar,India.Her research area includes software engineering.