Software Cost Estimation Using Data Mining: Review

[1]Miss Sumera w.Ahmad, [2]Dr.G.R.Bamnote

[1]Research Scholar, [2]Professor
Department of Computer Science & Engg
P.R.M.I.T & R, Badnera
Amravati, India

***Abstract*— Software cost estimation has become a great matter of concerns in the software industry. Accurate cost estimation in software development is very important for every kind of project. If the project is not estimated in the proper way it result the cost of the project very high extending the original cost. This needs the estimation accurate. This paper reviews all the different approaches that are used for software cost estimation using data mining. The different approaches help us to identify the challenges and scope present in software cost estimation using data mining.**

***Keywords*— Software estimation, COCOMOII, Data mining. Software engineering.**

I. INTRODUCTION

Software cost estimation is all about how long how many people are required to complete the software project. Project estimation starts at the proposal state and continues throughout the life time of the project. The estimation include size estimation, efforts estimation, & developing initial projects schedule and finally estimating overall cost of project .Accurate cost estimation is very important for every kind of project if the project is not estimated in proper way it result into high cost of project. This review paper will include the two different field i.e software estimation and data mining. For software cost estimation there are various standard models available such as SEER SEM, PUTMUN's, COCOMOI 81 and COCOMO II, among these COCOMO II is mostly and widely used for software cost estimation. Estimation process of COCOMO needs the cost drivers and scale factors values for accurate estimation. Data mining helps us to classify the past project data and generate the valuable information. These information is applied in cost estimation model COCOMO II and will generate the accurate estimation on the bases of past project data. For analyzing the data the analytical tool is Data mining. It allows users to analyze data from many different dimensions or angles categorize it and summarize the relationships identified there are various data mining tools available to identify the important and most common cost drivers of COCOMO II model that are used to generate the estimate of a project. Cost drivers are multiplicative factors which determine the effort required to complete the software project. In the analogy estimation models the cost drivers are the base of cost estimation models. They estimate the new project with comparing the past project data or cost drivers and set the value of cost drivers in the new projects.

II. RELATED WORK

A. Software Cost Estimation

Observing from the last decades software project cost estimation in software engineering is an important subject .Software project usually does not fail during implementation and mostly the project failure is related to the planning and estimation phase. Author Dharmesh & Mahesh [1] have perform analysis of different ANN's and compared the result of various ANN models of effort estimation .Author Adiano & Oliveria [2] has compared SVR (support Vector Regression) and Radial Bias Feed forward Network for software cost estimation .Lefly & sheppord [3] had improve the software cost estimation on public data sets by applying genetic programming and had achieved a great successes. Author Barbara & Magne [4] in their paper had made a comparison between evidence based software engineering with evidence based medicine. Author Chen[5] had proposed the cost models should be data pruned by experimenting after data collection and before model building .Author A.F.Shete [6] had used genetic algorithms on COCOMO model for estimation of projects.Galooroth & Evans [7] had performed intensive search between 2100 internet sites and had evaluated 500 reasons for software failure. Author Magne & Sheppard [8] has identified 304 software cost estimation papers in 76 journals in their review and classified them according to research topic,estimation approach,research approach ,study context and data sets. Author K.Vinay Kumar & V.Ravi [9] used a wavelet neural network for prediction of software cost estimation. Prasad & Reddy [12] has design a model for software cost estimation which has used a Multi Objective Particle Swarm Optimization algorithm. Its observation provides that the model provide with more good results when a comparison is made with the standard COCOMO model. [13] Author Pahariyaa & Ravia with more co authors have design a new computational intelligence sequential hybrid architectures which uses a Genetic Programming and Group Method of Data Handling and also include the recurrent architecture for Genetic Programming for better and accurate software cost estimation.

B. Data Mining

The Data mining at its depth means the transformation of huge amounts of data into purposeful patterns and rules. In general data mining tasks is divided into the following two categories: descriptive and predictive. Descriptive mining means cauterizing the data on the bases of general properties of data into databases. Predictive mining tasks deals with performing inference on current data for prediction. In different cases users may not known with the required kinds of patterns in their data may be interested and hence may like to search for

several different kinds of patterns in parallel. Thus data mining has an importance to have a data mining application system that can mine various kinds of patterns to accommodate different user expectations or applications. In Future the data mining systems should be able to discover patterns at various levels that is different levels of abstraction. Data mining systems must also permit users to specify clues to guide or center the search for interesting patterns because some patterns may not hold for all of the data in the database a measure of certainty or beleviable is usually associated with each discovered pattern. Now a day's data mining is use in every field of applications such as medical, banking, insurances,, education etc. It has been also applied to software artifacts within the realm of software engineering that is mining software repositories. Author Lum [10] had proposed a tool named 21st century effort estimation tool for new software cost estimation model using data mining technique. The accuracy of this model has been validated internally through leave one cross out validation. Author Katholieke [11] Techniques inducing tree/rule-based models like M5 and CART, Linear models such as various types of linear regression, nonlinear models\ (MARS, multilayered perceptron neural networks, radial basis function networks, and least squares support vector machines), and estimation techniques that do not explicitly induce a model .

III. CHALLENGES AND SCOPE

The paper propose to find out the common cost factors and scale drivers from the existing model of software cost estimation i.e COCOMO II model. And to estimate the accurate cost of the project with the help of past project data whose cost or effort is known. By using data mining algorithms and machine learning techniques into models of software cost estimation which will improve the performance of software cost estimation.

The Scale Drivers

In the COCOMO II model, some of the most important factors contributing to a project's duration and cost are the scale drivers. Scale drivers are set to describe the software project. Thus these Scale Drivers determine the exponent used in the Effort Equation. There are five scale driver used in the cocomo model and each cost driver play an important role in the estimation. The 5 Scale Drivers are Development Flexibility Precedentedness, Architecture/Risk Resolution, , Process Maturity, Team Cohesion.

The Cost Drivers

COCOMO II has 17 cost drivers for estimation of project, development environment and team to set each cost driver. The most important and multiplicative drivers are the cost drivers factors that determine the effort required to complete the software project. For example, if a project is developing a software based on aviation dealing with life and death if people would set the Required Software Reliability (RELY) cost driver to very high. This rating corresponds to an effort multiplier of 1.26 means that the project will be requireing 26% more effort than a typical software project. In the COCOMO II model, the cost drivers divide in the four groups stated below. **Personnel Factors**: Analyst Capability, Programmer Capability, Applications Experience Platform Experience, Personnel Continuity, Use of Software Tools. **Product cost driver**: Required Software Reliability, Data Base Size, Required Reusability Documentation match to life-cycle needs etc. **Platform Factors**: Execution Time Constraint, Platform Volatility. **Project Factors**: Required Development Schedule, Multisite Development etc.

COCOMO II EFFORT EQUATION

The COCOMO II model evaluates its estimates of required effort E (measured in Person-Months ï¿½ PM) based primarily on estimation of the software project's size (as measured in thousands of SLOC, KSLOC): Effort = 2.94 * EAF * (KSLOC)E. Where the Effort Adjustment Factor(EAF) has been derived from the Cost Drivers E which is an exponent derived from the five Scale Drivers As an example, a project with all Nominal value of Cost Drivers and Scale Drivers in COCOMO II would have an EAF of 1.00 and exponent, E, of 1.0997. Assume that the software project consist of 9,000 source lines of code, COCOMO II estimates that 29.9 Person-Months of effort is required to complete it. Effort = 2.94 * (1.0) * (9)1.0997 = 29.9 PM that is Person-Months.

Proposed work

The related work in section II have provided the way of software cost estimation using different ways. Let's considering the COCOMO II model of software cost estimation because of its agile and robust nature. Finding out the common cost drivers and scale drivers which are used by COCOMO II model for determining the exponent used in the effort equations and cost drivers which are multiplicative factor to determine the effort required to complete project. Scale Drivers and cost Drivers of the past projects which are stored in the software repositories are extracted through data mining (algorithm). Not all the data mining techniques performed better than the traditional method of local calibration .The proposed algorithm for data mining is classification data mining algorithms which are best suited for the predictive data mining category. This will provide an accurate cost estimation of the project which is proposed from the past history whose effort is known.

Conclusion

For the improvement in the accuracy of software cost estimation using a data mining and machine learning into the existing software estimation model i.e. COCOMO II. It will encourage practitioner's to shift from archaic estimation method and to select estimation tools that incorporate less risk and less uncertainty into software cost estimation.

References

[1] **[Dharmesh , Mahesh 1997]** Dharmesh Santani ,Mahesh Bundele , Poonam Rijwani , "Artificial Neural Networks for Software Effort Estimation: A Review" International Journal of Advances in Engineering Science and Technology. Volume 3. ISSN: 2319-1120.

[2] **[Adriano, Oliveira 2006]** Adriano L.I., Oliveira, "Estimation of software project effort with support vector regression". Neurocomputing, Volume 69, Issue 13, Pages 1749-1753.

[3] [**Lefley Shepperd 2005]** ,Martin Lefley , Martin J. Shepperd.,"Using Genetic Programming to Improve Software Effort Estimation Based on General Data Sets".Genetic and Evolutionary Computation GECCO Volume 2724 of the series 2003.

[4] **[Barbara et.al 2004]** Barbara A.Kitchenham, Tore Dyba, Magne Jorgensen, " Evidence Based Software Engineering" ,Proceedings of the 26th international conference on Software engineering ICSE 04 IEEE 2004.

[5] **[Chen et.al, 2005]** Chen, Z.; Menzies, T.; Port, D.; Boehm, B., "Finding The Right Data For Software Cost Modeling," Software, IEEE , vol.22, no.6, pp.38,46, Nov.-Dec. 2005.

[6] **[Sheta, 2006]** A.F.Sheta, "Estimation of the COCOMO Model Parameters Using Genetic Algorithms for NASA Software Projects", Journal of Computer Science, vol.2, pp. 118-123, 2006.

[7] **[Galorath, Evans 2006]** Galorath and Michael W. Evans "Successful Software Planning, Measurement and Control". Auerbach Publications, 2006, ISBN: 0849335930 2006

[8] [Jorgensen,Shepperd 2007]M. Jorgensen and M. Shepperd, "A Systematic Review of Software Development Cost Estimation Studies," in IEEE Transactions on Software Engineering, vol. 33, no. 1, pp. 33-53, Jan. 2007.

[9] **[Vinaykumar, Ravi , Rajkiran 2008]** K. Vinaykumar, V. Ravi, M. Carr and N. Rajkiran ."Software cost estimation using wavelet neural networks ."Journal of Systems and Software" Volume 81 Issue 11, Pages 1853-1867 November, 2008 Elsevier Science Inc. New York, NY, USA

[10] **[Lum et.al, 2008]** Karen T. Lum, Daniel R. Baker, and Jairus M. Hihn "The Effects of Data Mining Techniques on Software Cost Estimation" Engineering Management Conference, 2008. IEMC Europe 2008. IEEE International , vol., no., pp.1,5, 28-30 June 2008

[11] **[Katholieke et.al, 2012]** Dejaeger,K.,Verbeke, W., Martens,D, Baesens,B."Data Mining Techniques for Software Effort Estimation: A Comparative Study" Software Engineering, IEEE Transactions on , vol.38, no.2, pp.375,397, March-April 2012

[12] [**Reddy, Raju]** Satyan , A Reddy , Kvsvn Raju "An Improved Fuzzy Approach for COCOMO's Effort Estimation using Gaussian Membership Function" www.researchgate.net/.../42804622.2009

[13] **[Pahariyaa , Ravia, Carra, Vasua 2010]** J.S.Pahariyaa,b, V. Ravia,*, M. Carra and M. Vasua,b "Computational Intelligence Hybrids Applied to Software Cost Estimation" International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) Vol.2 (2010), pp.104-112 ISSN: 2150-7988.