Using classification Data Mining Techniques for Software Cost Estimation

Snehal A. Deshmukh, Prof. S. W. Ahmad

Abstract— Software cost estimation is the forecasting about the amount of effort required to make a software system and its duration. It is one of the basic project management processes carried out to support efficiently resource allocation activities. So, a lot of attention has been allocated to cost estimation. In required cost estimation estimated size and cost of the project with high accuracy is still vast challenge for projects managers. The accuracy of the cost estimation of software projects is necessary for the software companies. For the forecasting of software cost, it is important to select the correct software cost estimation techniques. Inaccurate cost estimation can be risky to an IT industry's economics.

Index Terms— Software Cost Estimation; Data Mining

1 INTRODUCTION

Software cost estimation

Software development cost estimation is vital for the professionals in IT industry . This critical task affects the firm's software investment decisions before starting for a contract or committing required resources to that project. Accurate software cost estimates are critical to both developers and customers. They can be used for generating request for proposals, scheduling, monitoring and control. A lot of software projects have been developed for accurate cost estimation. But it is difficult to say that there is any model that can give estimation close to the actual cost. Both over-estimation and under-estimation may lead to using the resources inefficiently, delaying the final software product delivery, unexpected increase in budget, or low quality of software projects. Thus no accurate decision can be made. The paper presents various methods for improving software cost estimation.

Data mining

Data mining is the process of exploration and analysis of large data, so that meaningful pattern and rules can be discovered.. The objective of data mining is to design and work efficiently with large data sets. Data mining is the component of wider process called knowledge discovery from database. Data Mining is the process of analysing data from different perspectives and summarizing the results as useful information.

The definition of data mining is closely related to another commonly used term knowledge discovery. Data mining can be stated as multi-step process which includes accessing and preparing data for a mining the data, data mining algorithm, analysing results and taking appropriate action. The data, which is discovered, can be stored in one or more operational databases. In data mining the data can be mined by passing various processes.

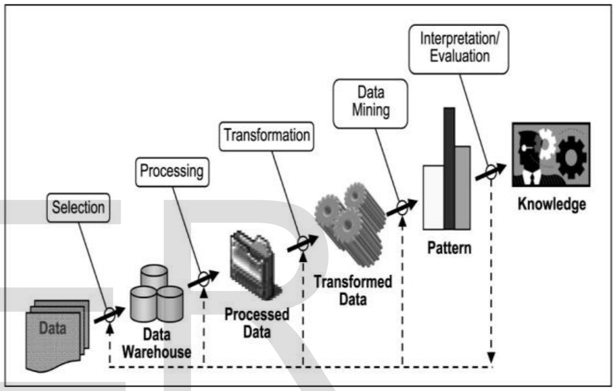

Fig1: Steps in data mining process

Data collected from multiple sources is integrated into a single data storage called as target data. Data relevant to the analysis is retrieved from the data collection. Then, it is pre-processed and transformed into an appropriate format suitable for mining. Data mining is a crucial step in which intelligent algorithm/techniques are applied to extract meaningful pattern or rules. Finally, those patterns and rules are interpreted to new or useful knowledge or information

2 RELATED WORK

Software Cost Estimation (SCE) is one of complex problems in software engineering area. So, a lot of attention has been allocated to solving this problem. Thus the correct estimate of the cost and effort required for software companies and company executives and companies is very significant. In required effort prediction, estimated size and cost of the project with high accuracy is still vast challenge for projects masters. In 1981, had been developed algorithmic models which called Constructive Cost Model (COCOMO) by Boehm. [2] These models were used for required effort prediction. Since, COCOMO 81 is not suitable model for responding new requirements in recent years, so COCOMO II model in 2000 published. COCOMO II provides tool for conducting empirical analysis of the model. [1]

- *Snehal A. Deshmukh is currently pursuing masters degree program in Comp.Science Department at PRMIT&R, Badnera.*
- *S. W. Ahmad is currently working as an Asst. Professor in Comp.Science Department at PRMIT&R, Badnera.*

Later, Vinaykumar [3] used wavelet neural networks for the prediction of software cost estimation. Unfortunately the accuracy of these models is not satisfactory so there is always a place for more accurate software cost estimation techniques. Lefley and Shepperd applied genetic programming to improve software cost estimation on public datasets with great success.

Jyoti Shivhare [15] in 2014 presented a paper and described a 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[14] 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. Neha-Saini in 2014 evaluated various machine learning techniques for software effort estimation like bagging, decision trees, decision tables, and 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 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 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 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 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. [6]

Menzies et al. proposed that feature subset selection should be regularly carried out in software cost estimation. The authors utilised a Wrapper linear regression and ranked the frequency with which each feature was selected by the algorithm to form groups of features with the same rankings. Features were then removed in ranked order. Using the best features the authors found that effort prediction results were always improved, even though for datasets with too many projects, the improvement was very small. Keung et al. proposed Analogy-X to select the best set of features and validate the appropriateness of estimation by analogy using Mantel's Correlation. The authors applied a method similar to stepwise regression analysis to perform sensitivity analysis, detect significant relationships, extract features and locate abnormal data. However, the method's main limitation was its inability to handle categorical values. Li et al. discussed how most FSS methods in analogy-based estimations are implemented as Wrappers and taking into account the advantages of Filter approaches such as selecting more appropriate features than wrappers that merely optimise the error measure, they proposed a hybrid Wrapper and Filter algorithm. Their results showed that the proposed method was an even more effective feature selector that could overcome some of the limitations and computational costs of other techniques proposed in the field [4]. The paper by K.Smith, et.al. has discussed the influence of four task assignment factors, team size, concurrency, intensity, and fragmentation on the software effort. These four task assignment factors are not taken into consideration by COCOMO I and COCOMO II in predicting software development effort. The paper by Girish H. Subramanian, et.al. concluded that the adjustment variables i.e. software complexity, computer platform, and program type have a significant effect on software effort. Boehm and Mendes [2] suggested that relying on organization-specific datasets leads to poor software cost predictions due to the following problems: collection of data on previous projects from single organization could be too expensive; information on older projects may outdated and no longer be valid or appropriate due to the new technologies that organization is using; and difficulty to ensure consistency of the collected data.

3 PROPOSED WORK

A. Aim

The aim of this dissertation work is to identify the important cost drivers in the past project data with the help of data mining classificaton techniques .Cost drivers are multiplicative factors of cost estimation model that determine the effort required to complete software project. In the analogy estimation models, the cost drivers are the base of cost estimation models. They estimate the new project with compare the past project data or cost drivers and set the value of cost drivers in the new projects.

B. Problem statement

Following problems often occurs due to inaccurate effort-estimation.

- Software project personnel have no firm basis for telling a manager, customer, or salesperson that their proposed budget and schedule are unrealistic.

• Analysts have no idea for making realistic hardware-software trade-off analysis. This often leads to inaccurate effort-estimation.

• Inaccurate effort-estimation leaves managers with no way to tell whether or not the software is proceeding according to plan.

C. Objectives

- To identify the cost drivers and cost factors of Cocomo2 model.
- To apply data mining classification techniques for software cost estimation.
- To determine what resources to commit to the project and how well these resources will be used as projects can be easier to manage and control when resources are better matched to real needs.
- To improve the overall business plan of a software organization
- To classify and prioritize development projects with respect to an overall business plan.
- To provide more accurate estimated software project to the customers.

3 CONCLUSION

In the above research work, an overview of different types of software cost estimation methods and also the advantages and disadvantages of these methods is provided along with reasons that cause inaccurate estimation. To have a reliable estimate, we must understand relationships between software projects and their attributes. We must develop effective ways of measuring software complexity and the cost estimation process needs to be thoroughly planned. All estimation methods are specific for some specific type of projects. It is very difficult to decide which method is better than to all other methods because every has its own significance. To understand their advantages and disadvantages is very important when you want to estimate your projects. In recent year research, researchers worked with the software engineering like data mining and machine learning techniques for improving the accuracy of software cost estimation process. The future work includes study of new software cost estimation methods and models that help us to easily understand the software cost estimation process.

REFERENCES

[1] Narendra Sharma, Ratnesh Litoriya "Incorporating Data Mining Techniques on Software Cost Estimation: Validation and Improvement" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 3, March 2012)

[2] Boehm, B., Clark, B., Horowitz, E., Madachy, R., Shelby, R., and Westland, C. "Cost models for future software life cycle process: COCOMO 2.0". In Annals of Software Engineering Special Volume on Software Process and Product Measurement. J. D. Arther and S. M. Henry, Eds., vol. 1, pp. 45–60, J.C. Baltzer AG, Science Publishers, Amsterdam, The Netherlands, 1995

[3] Vinaykumar, K., Ravi, V., Carr, M. and Rajkiran, N. "Software cost estimation using
wavelet neural networks," Journal of Systems and Software, 2008, pp. 1853-1867.

[4] Sonam Bhatia, Varinder Kaur Attri "Implementing Decision Tree for SoftwareDevelopment Effort Estimation of Software Project" International Journal of Innovative Research in Computerand Communication EngineeringVol. 3, Issue 5, May 2015

[5] Efi Papatheocharous1, Harris Papadopoulos2 and Andreas S. Andreou3, " Feature Subset Selection for Software Cost Modelling and Estimation"

[6] 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

[7] Mohita Sharma*1, Neha Fotedar2 Software Effort Estimation with Data Mining Techniques- A Review INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY 3(3): March, 2014]

[8] Ms. K. Gayathiri 1, Dr. T. Nalini 2 ,Dr. V. Khanaa " Data mining techniques for software effort estimation to improve cost efficiency" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 4 April, 2013 Page No. 1215-1220

[9] 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

[10] Muhammad Waseem Khan, Imran Qureshi " Neural Network based Software Effort Estimation: A Survey" Int. J. Advanced Networking and Applications Volume: 05, Issue: 04, Pages:1990-1995 (2014) ISSN : 0975-0290

[11] Hasan Al –Sakran "Efficient approach to develop software Cost estimation model using case-based reasoning and agent technology" Journal of Theoretical and Applied Information Technology 31st May 2014. Vol. 63 No.3

[12] Saiqa Aleem1, Luiz Fernando Capretz1, and Faheem Ahmed2 "Benchmarking Machine Learning Techniques For Software Defect Detection" International Journal of Software Engineering & Applications (IJSEA), Vol.6, No.3, May 2015

[13] T. C. Sharma & M. Jain (2013) "WEKA approach for comparative study of classification algorithm", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 4, 7 pages

[14] 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

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