Performance Evaluation of Software Effort Estimation using Fuzzy Analogy based on Complexity

S. Malathi
Research Scholar
Sathyabama University
Chennai, INDIA.

S. Sridhar
Research Supervisor
Sathyabama University
Chennai, INDIA.

ABSTRACT

Rapid industrialization in the past few decades has necessitated the ever increasing demand for newer technologies leading to the dramatic development of sophisticated software for cost estimation and is expected to grow manifold in the forthcoming years. The improper understanding of software requirements has often resulted in inaccurate cost estimation. In analogy concept, there is deficiency in handling the datasets containing categorical variables though there are innumerable methods to estimate the cost. The proposed fuzzy analogy method is a new approach based on reasoning by analogy for handling both numerical and categorical variables where the uncertainty and imprecision solution is ascertained by studying the behaviour pattern of linguistic values utilized in the software projects. The performance of linguistic values in fuzzy sets has improved in the proposed method. The performance of this method analyzed using Mean Absolute Relative Error (MARE) and Variance Absolute Relative Error (VARE) criteria indicates that the fuzzy analogy outperforms other techniques in terms of both quality and accuracy of the results in software cost estimation.

Keywords

Fuzzy analogy, Datasets, Cost estimation, Categorical variables, Linguistic values.

1. INTRODUCTION

Software cost estimation process is a set of procedures that an organization utilizes to arrive at a software cost estimate. Organizations have different software processes depending on the type of software that have been identified and developed. Evaluation by the existing methods gives vague and ambiguous results for both small and large organizations regardless of its immense significance in software cost estimation. Recent publications propose an integrated method of analogy and fuzzy for improving the performance of effort during the initial stages of cost estimation [1]. In [2], fuzzy logic using fuzzy numbers and function point analysis are utilized to represent the linguistic variables in order to reduce the imprecision and uncertainty problem. However, it is not suited for large datasets.

Focusing on imprecision and uncertainty alone is not adequate. There should be an integrated approach, incorporating different methodologies, to overcome the imprecision and uncertainty problem as well as to handle the categorical variables. This paper is intended for software engineers and others to predict an accurate estimation of the categorical and numerical data. On evaluation, it is found that for the datasets used in this paper, the fuzzy analogy method provides significantly better estimate than other comparable methods which are in vogue. Moreover, the success of this technique has thrown light on the vast scope for future research.

The rest of this paper has been dealt as follows. Section 2 presents the related work and Section 3 describes the implementation of fuzzy analogy method. In Section 4, results of the experiments with the existing datasets have been dealt while Section 5 discusses the conclusion and future research required in this approach.

2. RELATED WORK

Accurate effort estimation is an irreplaceable tool for an effective software project management. Extensive research has been carried out for the past few decades and several techniques have been evolved to predict estimation of effort accurately [3]. Recent literature has established that linear regression model [4] is often applied for cost estimation. However, there are limitations for the user in regression models while obliterating normal errors and moderate outliers [5].

Estimation by analogy [6] is the method where the proposed project is compared with earlier projects of similar nature with ample information in project development. The advantage of analogy method [7] in comparison to other methods is that analogy is based on actual experience. However, it is not very effective [8] due to non existence of similar projects and the precision of accessible historical data. In [9], genetic algorithm is applied in analogy to minimize the time involved while selecting the historic projects. Liu et al [10] has proposed a statistical framework for the elimination of noise and achieve enhancement of results in analogy method.

Despite the reality that analogy based method is one of the renowned methods of cost prediction, fuzzy technique using fuzzy numbers is employed to improve the accuracy in many areas like Control Engineering [11]. Fuzzy Logic has been the cynosure of recent important research investigations. During the early nineties, fuzzy logic has assumed its significance in terms of its theoretical approach [12]. Other techniques like fuzzy systems are resourceful in evaluating the effort using two-sided Gaussian membership function [13], by assigning the accurate degree of compatibility.

It is perceived that the accuracy of frequently used software models like COCOMO is only satisfactory [14]. Several studies have been carried out to update the COCOMO II [15] by filtering the cost drivers for future estimation. Newer approaches like min-max approach [16] improve the accuracy in COCOMO model. A new concept was introduced to outline the failure of effort estimation due to misleading information [17] on account of irrelevant data. However, later results have confirmed that instance selection [18] and retrieval are automatically done to reduce the unrelated data but still a unified and integrated approach is invariably essential to avoid any perception of errors.

3. PROPOSED WORK

3.1 Effort Estimation

In this section, a new method has been proposed to select the categorical variables and overcome the uncertainty problem in the software estimation process using default packages of JAVA. The proposed method combines the analogy concept with the fuzzy method. Analogy-based estimation has motivated considerable research in recent years. However, none or even a very few have yet to deal software cost estimation with categorical data. The key activities for estimating software project effort by analogy are the identification of a candidate software project as a new case, the retrieval of similar software projects from a project's repository and the reuse of knowledge derived from previous software projects to generate an estimate for the candidate software project.

Fuzzy logic is based on human behavior and reasoning. It has an affinity with fuzzy set theory and applied in situations where decision making is difficult. A Fuzzy Set can be defined as an extension of classical set theory by assigning a value for an individual in the universe between the two boundaries that is represented by a membership function.

$$A = \int_{x} \mu_{A}(x)/x \qquad (1)$$

Where x is an element in X and $\mu_A(x)$ is a membership function.

3.2 Fuzzy Analogy

Fuzzy Analogy is the fuzzification of the classical analogy procedure and composed of three steps: case(s) identification, retrieval of similar cases and case adaptation. Each step is a fuzzification of its equivalent in the classical analogy-based estimation procedure. In the following sub-sections, each fuzzified step has been discussed with more details.

3.2.1 Identification of Cases

The aim of this step is to characterize all software projects by a set of attributes. Selection of attributes that accurately describe software projects is a complex task in the analogy-based procedure. The primary objective of this study is to estimate the software project effort. Consequently, the attributes must be relevant for the effort estimation task. The framework is shown in figure.1

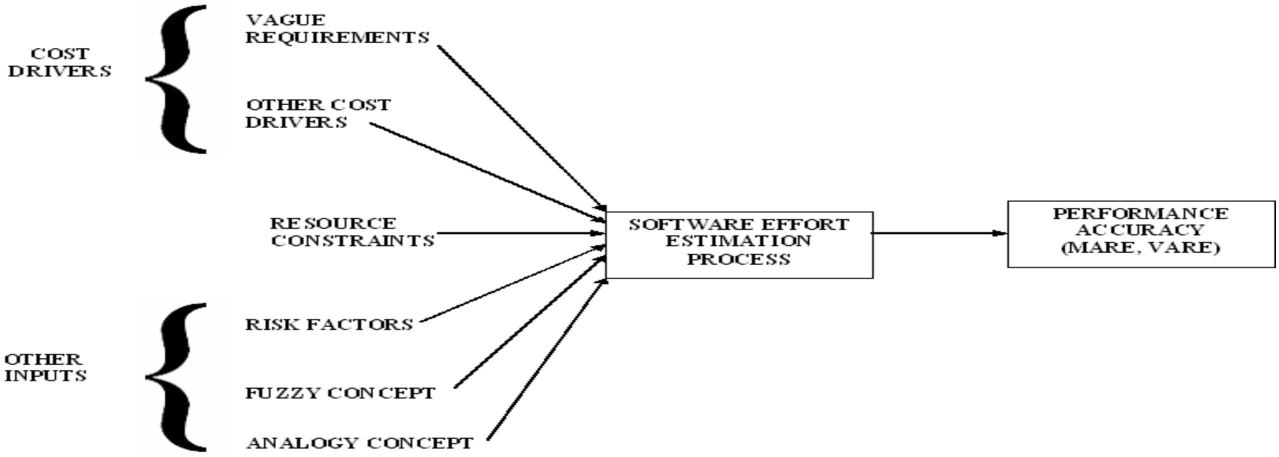

Fig 1: Framework of Fuzzy Analogy

The problem is to detect the attributes exhibiting a significant relationship with the effort in a given environment. The attributes are represented by fuzzy sets. A fuzzy set is a set with a graded membership function, μ, in the real interval [0, 1].

In the case of numerical value x_0, its fuzzification will be done by the membership function which takes the value of 1 when x is equal to x_0 and 0 otherwise. For categorical values, M attributes are considered and for each attribute M_j, a measure with linguistic values is defined (A_k^j).

Each linguistic value A_k^j is represented by a fuzzy set with a membership function ($\mu_{A_k^j}$).

The representation of the categorical variables such as 'very low' and 'low' is based on how humans interpret these values and consequently it allows dealing with imprecision and uncertainty in the steps to be followed. The framework is built upon an existing cost estimation model, namely, COCOMO. The facts were analyzed to determine the formulas that are the best fit to the observations. This formula is related to the size of the system and product, project and team factors for the effort to develop the system. In COCOMO, effort is expressed as Person Months (PM). Cost drivers have up to six levels of rating: Very Low, Low, Nominal, High, Very High, and Extra High [19].

3.2.2 Retrieval of Cases

This step is based on the preferred choice of a software project similarity measure. The selection is obviously very critical since it will influence the type of analogies or similar cases extracted from the data set. The similarity of two software projects, which are described and characterized by a set of attributes, is often evaluated by measuring the distance between these two projects through their sets of attributes. Thus, two projects are considered dissimilar if the differences between their respective sets of attributes are clear and obvious.

These measures assess the overall similarity of two projects P_1 and P_2, $d(P_1, P_2)$ by combining all the individual similarities of P_1 and P_2 associated with the various linguistic variables V_j describing the project P_1 and P_2, $d_{V_j}(P_1, P_2)$. After an axiomatic validation of some proposed candidate measures for the individual distances $d_{V_j}(P_1, P_2)$, two measures have been retained [20].

$$d_{V_j}(P_1,P_2) = \begin{cases} \max\limits_{\text{k}} \ \min(\mu_{A_k^j}(P_1), \mu_{A_k^j}(P_2)) \\ \max - \min \ \ aggregation \\ \sum\limits_{k} \mu_{A_k^j}(P_1) \times \mu_{A_k^j}(P_2) \\ sum - product \ \ aggregation \end{cases} \quad (2)$$

Where A_k^j are the fuzzy sets associated with V_j and $\mu_{A_k^j}$ are the membership functions representing fuzzy sets A_k^j. Scale factors (SF) are understanding product objectives, flexibility, team coherence etc., Effort multipliers (EF) are software reliability, database size, reusability, complexity etc. The imprecision of the cost drivers significantly affects the accuracy of the effort estimates which are derived from effort estimation models. Since the imprecision of software effort drivers cannot be overlooked, a fuzzy model gains advantage in verifying the cost drivers by adopting fuzzy sets.

$$Effort = A * (SIZE)^{B + 0.01 * \sum_{I=1}^{N} SF_i} * \prod_{i=1}^{N} EM_i \quad (3)$$

Where A and B are constants, SF is the scale factor and EM is effort multipliers. By using the above formula the effort is estimated. The cost drivers are fuzzified using triangular and trapezoidal fuzzy sets for each linguistic value such as very low, low, nominal, high etc. as applicable to each cost driver. Rules are developed with cost driver in the antecedent part and corresponding effort multiplier in the consequent part. The defuzzified value for each of the effort multiplier is obtained from individual Fuzzy Inference Systems after matching, inference aggregation and subsequent Defuzzification. Total Effort is obtained after multiplying them together. The high values for the cost drivers lead an effort estimate that is more than three times the initial estimate, whereas low values reduce the estimate to about one third of the original.

3.2.3 Case Adaptation

The objective of this step is to derive an estimate for the new project by using the known effort values of similar projects. There are two issues that have to be addressed, (i) the choice of how many similar projects should be used in the adaptation, and (ii) how to adapt the chosen analogies in order to generate an estimate for the new project. In the available literature, it can be clearly noticed that there is no definite rule to guide the choice of the number of analogies. Fixing the number of analogies for the case adaptation step is considered here neither as a requirement nor as a constraint.

4. EXPERIMENTAL RESULTS

In the experimental study, 30% of the data from the NASA 93 dataset [21] is chosen from different NASA centers to measure the estimated effort with the actual effort. From the results, it was found that the estimated effort is very low compared to the existing actual effort. The comparison of the estimated effort using COCOMO, fuzzy GBellMF and Fuzzy Analogy are shown in Table 1.

Table 1 Comparison of Estimated Effort from Various Methods

Project ID	Actual Effort	Estimated Effort		
		COCOMO	*Fuzzy GBellMF*	*Fuzzy Analogy*
21	60	50.60	80.99	802.01
45	400	400.00	483.11	766.93
46	2400	2400.00	3212.00	1201.7
47	420	436.90	483.07	13.90
53	750	703.06	743.06	748.00
54	2120	2120.00	1137.81	633.14
62	2468	2356.00	2164.00	1249.8
75	600	600.00	586.92	703.5
82	480	478.01	446.20	695.17
83	599	632.16	650.12	133.4
85	4148.2	8432.62	4027.18	1020.71
86	1772.5	1239.60	1624.30	1043.58
87	1645.5	1546.20	986.20	1094.62
88	1924.5	1627.18	1889.32	1594.46
93	38	31.90	25.34	2.67

The estimated effort using the proposed method and the actual effort is represented in Figure.2 and the comparison of various methods is shown in Figure 3.

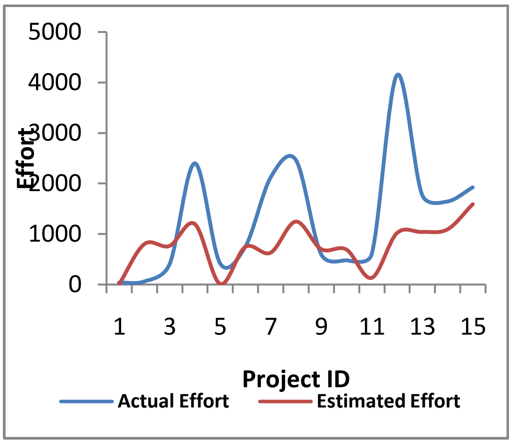

Fig.2: Estimated Effort using Fuzzy Analogy and Actual Effort

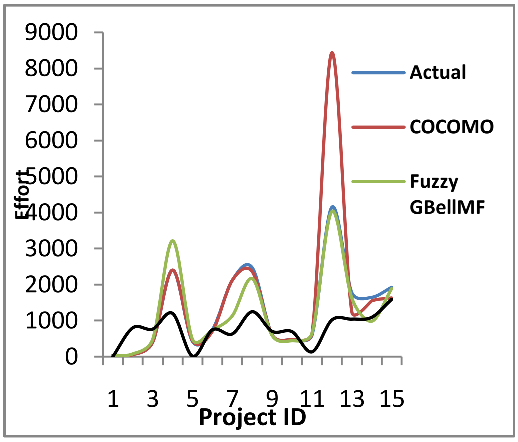

Fig.3: Project ID versus Comparison of various methods

4.1 Performance Criterion for Assessment of Software Effort

Various researchers have used different error measurements[22]. One of the criteria for the evaluation of cost estimation is the Mean Absolute Relative Error (MARE).

$$MARE = \frac{MRE}{n} \qquad (4)$$

Where ‘n ‘is the number of projects in a dataset and

$$MRE = \sum_{i=1}^{n} (|act_i - est_i |) / |act_i | \qquad (5)$$

Where est_i is the estimated effort from the model and act_i is the actual effort, and n is the number of projects in the model. Another criterion of measurement is the Variance Absolute Relative Error (VARE).

$$VARE = Var\left[\frac{abs(|est_i - act_i|)}{act_i}\right] * 100 \qquad (6)$$

Thorough evaluation of NASA 93 dataset is demonstrated using the proposed method and compared with the existing COCOMO and Fuzzy GBellMF based on MARE and VARE measure [19] which is tabulated in Table 2 and represented in Figure 4 and Figure 5.

Table 2 Comparison of Various Methods

Methods	VARE	MARE
Fuzzy Analogy	2.96	1.426
Fuzzy GBellMF	32.59	23.78
COCOMO	47.22	46.89

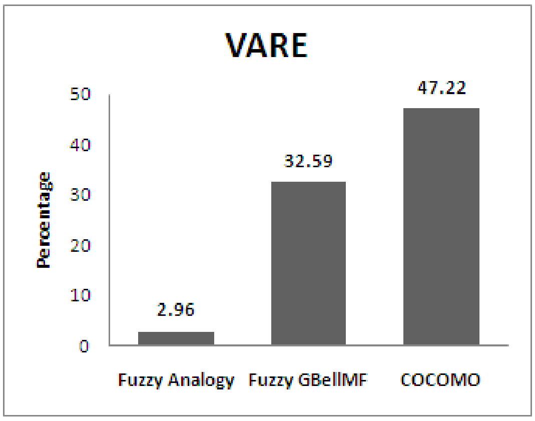

Fig.4: Comparison of VARE against various methods

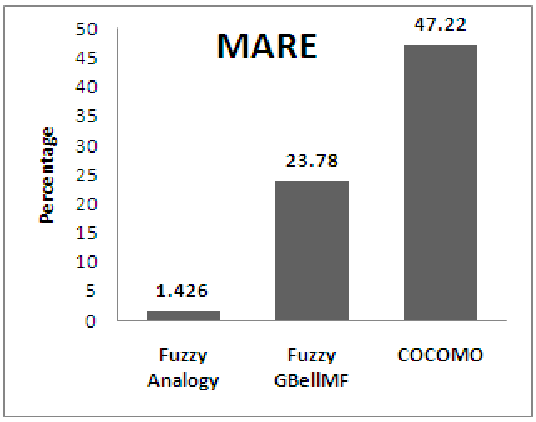

Fig.5: Comparison of MARE against various methods

From the performance criterion, it is observed that the proposed method shows significant results. Overall findings indicate that there is a tendency to improve the MARE value if the dataset is large with more number of categorical variables. Perhaps, the main finding that can be drawn on this point is that the larger the dataset, the outcome is likely to be more reliable. This point needs further investigation and has to be corroborated by checking with smaller datasets. Further, empirical investigation is also necessary to ensure the validity of the proposed approach on the other numerical datasets

5. CONCLUSION

In this paper, a new approach has been proposed to estimate the software project effort. This approach is based on reasoning by analogy, fuzzy logic and linguistic quantifiers. Such an approach can be used when the software projects are described by categorical or numerical data, through there is no reliable method to handle the categorical variables. Thus, the approach improves the classical analogy procedure which does not take into account of the categorical data. In the fuzzy analogy approach, both categorical and numerical data are represented by fuzzy sets. The advantage of this method is that it can handle correctly the imprecision and the uncertainty of the software project. The findings suggest that an effective strategy in increasing the accuracy of the software cost estimation is to combine the fuzzy logic with the analogy concepts. However, it has to be substantiated with expert opinion which can be integrated to improve the accuracy and reliability of this procedure. In future, the best way to handle the uncertainty and the categorical data is to let the project managers to find out the individual measurements of uncertainty during data collection.

6. REFERENCES

[1] Mohammad Azzeh,Daniel Neagu,Peter I.Cowling, 2011. “Analogy based software effort estimation using fuzzy numbers”, The journal of systems and software 84 270-284.

[2] Yinhuan Zheng, Beizhan Wang, Yilong Zheng, Liang Shi,2009. “Estimation of software projects effort based on function point,” Proceedings of 4th International Conference on Computer Science & Education.

[3] Jacky Keung, 2009. “Software Development Cost Estimation using Analogy: A Review, “Australian Software Engineering Conference, pp. 327-336.

[4] Berlin.S., Raz.T., Glezer.G., Zviran.M., 2009. Comparison of estimation methods of cost and duration in IT projects. Information and Software Technology 51 (4), 738-748.

[5] Kichenham.B.A., Mendes.E., 2009. Why comparative effort prediction studies are invalid. In: PROMISE09’ Proceedings of the 5th International Conference on Predictor Models in Software Engineering.

[6] Wai, J., B. Keung, A. Kitchenham and D.R. Jeffery, 2008.” Analogy-X: Providing statistical inference to analogy-based software cost estimation “, IEEE Transactions Software Eng., Vol. 34, pp. 471-484.

[7] Jianfeng Wen, Shixian Li, Linyan Tang , 2009.“Improve Analogy-Based Software Effort Estimation using Principal Components Analysis and Correlation Weighting,” 16th Asia-Pacific Software Engineering Conference.

[8] Joon-kil Lee, Ki-Tae Kwon,2009. ” Software Cost Estimation using SVR based on Immune Algorithm,” 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing.

[9] Y.Li, M. Xie, and T.Goh, 2009. “A study of project selection and feature weighting for analogy based software cost estimation, “Journal of systems and software, vol.82, pp.241-252.

[10] Q. Liu, W.Z. Qin, R. Mintram, M. Ross,2008. “ Evaluation of preliminary data analysis framework in software cost estimation based on ISBSG R9 data,” Software quality journal, 16(3): 411-458.

[11] Wei. S. –J., Chen. S. –M., 2009.A new approach for Fuzzy risk analysis based on similarity measures of generalized Fuzzy number. Journal of Expert Systems with Applications 36,589-598.

[12] Iman Attarzadeh and Siew Hock Ow,2010. “ A Novel Algorithmic Cost Estimation Model Based on Soft Computing Technique,” Journal of Computer Science 6 (2): 117-125.

[13] Liu, H. and L. YU,2005. “ Towards integrating feature selection algorithms for classification and clustering,” IEEE Transactions on Knowledge and Data Engineering, 17(4): 491-502.

[14] Majed Al Yahya, Rodina Ahmad, and Sai Lee, April 2010. “ Impact of CMMI Based Software Process Maturity on COCOMO II’s Effort Estimatiom, “ International Arab Journal of Information Technology, Vol. 7, No. 2.

[15] Huang X., Ho D., Ren J., and Capretz L,2007. “ Improving the COCOMO Model with a Neuro Fuzzy Approach,” Computer Journal of Applied Soft Computing Journal, Vol. 7, No. 3, pp. 29-40.

[16] H. S. Hota, Ramesh Pratap Singh, July 2011. “A min-max Approach for Improving the Accuracy of Effort Estimation of COCOMO,” International Journal of Soft Computing and Engineering (IJSCE), Vol.1, Issue 3.

[17] Magne Jørgensen and Stein Grimstad, Oct. 2011.“The Impact of Irrelevant and Misleading Information on Software Development Effort Estimates: A Randomized Controlled Field Experiment, “ IEEE Transactions on Software Engineering, vol. 37,No.5.

[18] Ekrem Kocaguneli, Tim Menzies, 2011.” How to Find Relevant Data for Effort Estimation?” International Symposium on Empirical Software Engineering and Measurement.

[19] Prasad Reddy P.V.G.D, Sudha K.R,Rama Sree P, 2011. “ Application of Fuzzy Logic Approach to Software Effort Estimation,” International Journal of Advanced Computer Science and Applications,Vol. 2, Issue 5.

[20] A.Idri and A. Abran, 2001. "Towards A Fuzzy Logic Based Measures For Software Project similarity", In Proc. of the 7th International Symposium on Software Metrics, England, pp.85-96.

[21] Sayyad Shirabad, J. and Menzies, T.J. 2005. The PROMISE Repository of Software Engineering Databases. School of Information Technology and Engineering University of Ottawa, Canada. Available: http://promise.site.uottawa.ca/SERepository

[22] Wei Lin Du, Danny Ho and Luiz Fernando Capretz, Oct.2010. "Improving Software Effort Estimation Using Neuro-Fuzzy Model with SEER-SEM", Global Journal of Computer Science and Technology, Vol. 10, No. 12, Pp. 52-64.