Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

Efficient estimation of effort using machine-learning technique for software cost

S. Malathi*[1] and S. Sridhar[2]
[1] *Research Scholar, Department of CSE, Sathyabama University, Chennai-600119, India*
[2] *Department of CSE & IT, Sathyabama University, Chennai-600 119, India*
malathi_raghu@hotmail.com*, dakyesyes @gmail.com[2]

Abstract

Several useful models have been developed by the software engineering community to elucidate the periodic growth of life cycle and calculate the effort of cost estimation in a precise manner. One of the commonly used machine learning techniques is the analogy method that cannot handle the categorical variables efficiently. In general, project attributes of cost estimation are often measured in terms of linguistic values. These imprecise values leads to analogous while explaining the process. The proposed fuzzy analogy method is a new approach based on reasoning by analogy using fuzzy logic for handling both numerical and categorical variables where the uncertainty and imprecision solution is also identified by the behavior of linguistic values utilized in the software projects. The performance of this method validates the results based on historical NASA dataset. The outcome of fuzzy analogy method is analyzed which indicates its improvement over the existing fuzzy logic methods.

Keywords: Analogy, Categorical variables, Cost estimation, Fuzzy logic, Linguistic.

Introduction

The software environment has evolved significantly in the last 30 years. To estimate software development effort, the use of the neural networks has been viewed with skepticism by majority of the cost estimation community. Even though neural networks have exposed their strengths in solving multifarious problems, their limitation of being 'black boxes' has limited their usage as a common practice for cost estimation (Pichai *et al.,* 2010). Some models carry a few advantageous features of the neuro-fuzzy approach, such as learning capability and excellent interpretability, while maintaining the qualities of the COCOMO model (Ahmeda & Muzaffar, 2009).

Estimation by analogy is simple and flexible, compared to algorithmic models. Analogy technique is applied effectively even for local data which is not supported by algorithmic models (Keung, 2008; Ekrem Kocaguneli *et al.,* 2010). It can be used for both qualitative and quantitative data, reflecting closer types of datasets found in real life. Analogy based estimation has the potential to mitigate the effect of outliers in a historical data set, since estimation by analogy does not rely on calibrating a single model to suit all the projects. Unfortunately, it is difficult to assess the preliminary estimation as the available information about the historic project data during early stages is not sufficient (Hasan Al-Sakran, 2006). The proposed method effectively estimates the software effort using analogy technique with the classical fuzzy approach.

Several researchers have carried out researches in the field of effort estimation for the software projects using various techniques (Jorgensen & Shepperd, 2007). A few of the significant researches have been highlighted here for iris recognition. Fuzzy logic has been applied to the COCOMO using membership.

Functions such as Symmetrical Triangles and Trapezoidal Membership Function (TMF) to signify the cost drivers. The limitation of the latter function is that a few attributes were assigned the maximum degree of compatibility instead of lower degree. To overcome this drawback (Satyananda Reddy & Raju, 2009) if proposed the usage of Gaussian Membership Function (GMF) for the cost drivers by studying the behaviour of COCOMO cost drivers. Kazemifard *et al.* (2011) used a multi agent system for handling the characteristics of the team members in fuzzy system. There are many studies that utilized the fuzzy systems to deal with the ambiguous and linguistic inputs of software cost estimation (Iman Attarzadeh & Siew Hock Ow, 2010).

Wei Lin Du *et al.,* (2010) followed an approach combining the neuro-fuzzy technique and the SEER-SEM effort estimation algorithm. The continuous rating values and linguistic values are the inputs of the proposed model for avoiding the deviation in estimation among similar projects. The performance of the proposed model has been improved by designing and evaluated with data from published historical projects. The evaluation results indicate that the estimation with the proposed fuzzy model containing analogy reasoning produce better results in comparison with the existing estimated results that uses feature selection algorithm.

Proposed work

Effort estimation

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.

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

Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

Where x is an element in X and $\mu_A(x)$ is a membership function. A Fuzzy set is characterized by a membership function that has grades between the interval [0, 1] called grade membership function. There are different types of membership function, namely, triangular, trapezoidal, Gaussian etc.

Fuzzy analogy

Fuzzification of classical analogy procedure is Fuzzy analogy. It comprises of three steps, 1) Identification of cases, 2) Retrieval of similar cases and 3) Case adaptation. Each step is the fuzzification of its equivalent classical analogy procedure.

Step 1: Identification of cases

The goal of this step is the characterization of all software projects by a set of attributes. Selecting attributes, which will describe software projects, is a complex task in the analogy procedure. Consequently, the attributes must be relevant for the effort estimation task. The objective of the proposed Fuzzy Analogy approach is to deal with categorical data. So, in the identification step, each software project is described by a set of selected attributes which can be measured by numerical or categorical values. These values will be represented by fuzzy sets.

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 ($\mu_{A_k^j}$).

It is preferable that these fuzzy sets satisfy the normal condition. The use of fuzzy sets to represent categorical data, such as 'very low' and 'low', is similar to how humans interpret these values and consequently it allows dealing with imprecision and uncertainty in the case identification step.

Step 2: Retrieval of Similar Cases

This step is based on the choice of software project similarity measure. 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 (Idri & Abran, 2001).

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

Where A_k^j the fuzzy sets are associated with V_j and $\mu_{A_k^j}$ are the membership functions representing fuzzy sets A_k^j. The effort is estimated using the formula:

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

Where A and B are the constants and d is the distance. Rules are developed with the cost driver in the antecedent part and the corresponding effort multipliers (EM) 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. This highlights the vast differences between different types of projects and the difficulties of transferring experience from one application domain to another.

Step 3: Case adaptation

The objective of this step is to derive an estimate for the new project by using the know effort values of similar projects. We are not convinced in fixing the number of analogies in this step. In our proposed method, all the projects in a dataset are used to derive the new project estimate. 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.

Experimental results

This section explains the accuracy of effort estimation by the proposed work as well as the performance against other methods. The standard datasets are chosen from the available software engineering public domain as follows. In this method, NASA 93 (Sayyad Shirabad &

Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

Menzies, 2005) was selected consisting of 93 projects in various programming languages. The implementation is done using the default packages of JAVA net beans. The estimated values of the proposed Fuzzy Analogy method for NASA 93 dataset is compared with the existing fuzzy method using triangular membership function and the outcomes are tabulated in Table 1.

Table 1. Comparison of proposed effort with the existing and actual effort

Pjt.ID	Actual Effort	Estimated Effort	
		Fuzzy Method	Fuzzy Analogy Method
1	36	45.31	34.14
2	42	32.37	36.48
3	42	43.83	36.60
4	50	60.74	43.625
5	60	80.99	52.125
6	120	113.77	103.122
7	72	94.04	61.794
8	192	172.59	163.201
9	239	268.04	203.150
10	300	354.73	254.512
11	352.8	354.73	324.538
12	420	483.07	355.993

Fig.1. Comparison of proposed effort with actual and fuzzy method

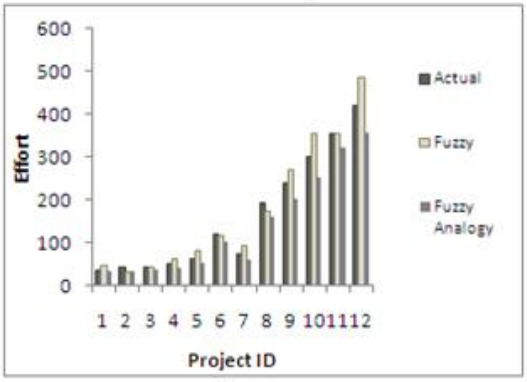
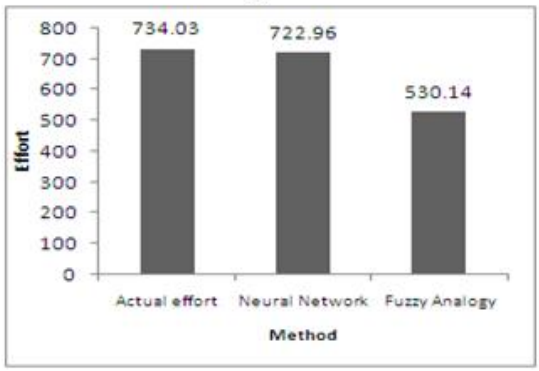


Fig. 2. Comparison of average effort with existing methods



The performance of the average effort in the proposed method is compared with the existing method (Prasad Reddy *et al.,* 2011) and found that the fuzzy analogy method is very efficientand shown in Fig.1. This method is also comaperd with the other machine learning techniques like neural network (Pichai *et al.,* 2010) and the values are predicted.From the comparative results shown in Fig. 2, it is predicted that the existing average effort using the selected features is very high compared to the proposed method.

Conclusion

An improved fuzzy analogy approach has been proposed in this paper to estimate the effort for the historical dataset. This approach is based on reasoning by analogy, fuzzy logic and linguistic variables. It can be applicable to numerical and categorical datasets and can handle the imprecision and uncertainty in an effective manner. The fuzzy analogy outperforms in estimating the effort; in future it can be optimized further by dealing many datasets in consideration.

References

1. Ahmeda MA and Muzaffar Z (2009) Handling imprecision and uncertainty in software development effort prediction: a type-2 fuzzy logic based framework. *Info. & Software Technol.* 51, 640-654.
2. Ch. Satyananda Reddy and KVSVN Raju (2009) An improved fuzzy approach for COCOMO's effort estimation using gaussian membership function. *J. Software.* 4(5), 452-459.
3. Ekrem Kocaguneli, Tim Menzies, Ayse Bener and Jacky W Keung (2010) Exploiting the essential assumptions of analogy-based effort estimation. *J. IEEE Trans. Soft. Eng.* 34(4), 471-484.
4. Hasan Al-Sakran (2006) Software cost estimation model based on integration of multi-agent and case-based reasoning. *J. Comput. Sci.* 2(3), 276-282.
5. Idri A and Abran A (2001) Towards A fuzzy logic based measures for software project similarity. *Proce. 7th Intl. Sym. Soft. Metrics.,* England, pp: 85-96.
6. Iman Attarzadeh and Siew Hock Ow (2010) Improving the accuracy of software cost estimation model based on a new fuzzy logic model. *World Appl. Sci. J.* 8(2), 177-184.
7. Jorgensen M and Shepperd M (2007) A systematic review of software development cost estimation studies. *IEEE Trans. Soft. Eng.* 33(1), 33-53.
8. Kazemifard M, Zaeri A, ghasem-ghaee N, Nematbakhsh MA and Mardukhi F (2011) Fuzzy emotional COCOMO II software cost estimation (FECSCE) using multi-agent systems. *Appli. Soft. Comput. Elsevier.* pp: 2260-2270,
9. Keung J (2008) Empirical evaluation of analogy-x for software cost estimation. *Proc.2nd ACM-IEEE Int. Sym. Empirical Eng. & Measurement.* NY, USA: ACM. pp: 294-296.
10. Pichai Jodpimai, Peraphon Sophatsathit and Chidchanok Lursinsap (2010) Estimating software effort with minimum features using neural functional approximation. ICCSA.
11. Prasad Reddy PVGD, Sudha KR and Rama Sree P (2011) Application of fuzzy logic approach to software effort estimation. *Int. J. Adv. Comput. Sci. & Appl.* 2(5), pp. 87-92.
12. Sayyad Shirabad J and Menzies TJ (2005) The PROMISE repository of software engineering databases. *School Info. Technol. & Eng. Univ.* Ottawa, Canada. Available: http://promise.site.uottawa.ca/SERepository.
13. Wei Lin Du, Danny Ho and Luiz Fernando Capretz (2010) Improving software effort estimation using neuro-fuzzy model with SEER-SEM. *Global J. Comput Sci. &Technol.* 10(12), 52-64.