Fuzzy Multiple Regression Model for Estimating Software Development Time

Venus Marza and Mir Ali Seyyedi
MSc student of Computer Engineering at Islamic Azad University of South Tehran Branch
Assistant Professor at the department of Computer Engineering, Islamic Azad University of South Tehran Branch
Corresponding author E-mail: Venus.Marza@gmail.com

Abstract: As software becomes more complex and its scope dramatically increase, the importance of research on developing methods for estimating software development time has perpetually increased, so accurate estimation is the main goal of software managers for reducing risks of projects. The purpose of this article is to introduce a new Fuzzy Multiple Regression approach, which has the higher accurate than other methods for estimating. Furthermore, we compare Fuzzy Multiple Regression model with Fuzzy Logic model & Multiple Regression model based on their accuracy.

Keywords: Fuzzy Logic (FL), Multiple Regression Model, McCabe Complexity, Dhama Coupling, Development Time

1. Introduction

Many study have already proposed models for size, effort, time and cost estimation. We just consider some of these studies: Regression analysis is a classical statistical technique for building estimation models. It is one of the most commonly used methods in econometric work. It is concerned with describing and evaluating the relationship between a dependent variable and one or more independent variables. The relationship is described as a model for estimating the dependent variable from independent variables. The model is built and evaluated through collecting sample data for these variables. This model was used first for estimating LOC of an information system (Kuan Tan et al. 2006).

Boehm was the first researcher to look at software from an economic point of view. Putnam also developed model known as SLIM, but both of COCOMO and SLIM are based on linear regression techniques (Moataz et al. 2005). Algorithmic models such as COCOMO, have failed to present suitable solutions that take into consideration technological advancements, because they are often unable to capture the complex set of relationships (e.g. the effect of each variable in a model to the overall prediction made using the model), they are not flexible enough to adapt to a new environment, and they can’t learn from their previous knowledge, also parametric models use a static predictive function for estimating (e.g. COCOMO use Effort = A \cdot Size ^ b for estimating Effort (Xia et al. 2005)).

Their inability contributed to exploring non parametric methods such as Fuzzy Logic, Soft computing which is a consortium of methodologies centering in Fuzzy Logic, Artificial Neural Networks and Evolutionary Computation. Research of MacDonell has evolved into development of a FULSOME (FUzzy Logic for SOftware MEtrics), to assist project managers in making predictions (Moataz et al. 2005, MacDonell 2003). Orginally, estimation was performed using only human expertise, but more recently, attention has turned to a variety of combining methods. Here we apply fuzzy concepts to regression model and compare their results with each other. The primary motivation of fuzzy set theory is the desire to build a formal quantitative structure capable of capturing the imprecision of human knowledge, that is, the manner in which knowledge is expressed in natural language. This theory seeks to bridge the gap that separates traditional mathematical models needed for physical systems, and the mental representation, generally imprecise, of such systems (Lima et al. 1999).

This paper is structured as follows. In section 2, multiple regression equation is considered and its result is shown in data subset, then in next section we apply fuzzy logic to the same data subset. In section 4 we introduce specific regression model with fuzzy concepts. In section 5, evaluation criteria are introduced for evaluating models, and in section 6, we apply these mentioned models to the same data subset for comparing them. Finally, conclusions are drawn in section 7.

2. Multiple Regression Equation

This model is the most common statistical technique for estimating. A linear equation with three independent variables (McCabe Complexity (MC), Dhama Coupling (DC), and physical Lines Of Code (LOC)) and a dependent one (Development Time (DT)) may be expressed as (Cuauhtemoc et al. 2005):

\[ DT = b_0 + b_1 \cdot MC + b_2 \cdot DC + b_3 \cdot LOC \]  

(1)

Where \( b_0, b_1, b_2 \) and \( b_3 \) is obtained by solving follow equations:
\[
\begin{align*}
\sum y &= b_0 + b_1(\sum x_1) + b_2(\sum x_2) + b_3(\sum x_3) \\
\sum x_1y &= b_0(\sum x_1) + b_1(\sum x_1^2) + b_2(\sum x_1x_2) + b_3(\sum x_1x_3) \\
\sum x_2y &= b_0(\sum x_2) + b_1(\sum x_1x_2) + b_2(\sum x_2^2) + b_3(\sum x_2x_3) \\
\sum x_3y &= b_0(\sum x_3) + b_1(\sum x_1x_3) + b_2(\sum x_2x_3) + b_3(\sum x_3^2)
\end{align*}
\] (2)

For simplify we used \(x_1\) as MC, \(x_2\) as DC, \(x_3\) as LOC and \(y\) as DT. By using data from Table 3 (Cuauhtemoc et al. 2005) and finding parameter \(b_0, b_1, b_2, b_3\) we can give following linear equation (Cuauhtemoc et al. 2005):

\[
DT' = 17.3097 + 2.06268^*MC - 32.9405^*DC - 0.049962^*LOC
\] (3)

The result of each module is organized in Table 4.

3. Estimating by Fuzzy Logic Rules

A fuzzy model like any other model provides mapping from input to output. For obtaining a fuzzy model first the verbal expert knowledge, based on the correlation (r) between pairs of their variables, is translated into if-then rules. Parameters of this structure, such as membership functions and weights of rules, can be tuned by using input and output data.

Correlation is the degree that indicates two sets how much are related to each other, and is defined as follows (Cuauhtemoc et al. 2005):

\[
r = \frac{n[\sum (X_iY_i) - (\sum X_i)(\sum Y_i)]}{\sqrt{n[\sum X_i^2 - (\sum X_i)^2][\sum Y_i^2 - (\sum Y_i)^2]}}
\] (4)

Correlation between development time as dependent variable and McCabe complexity, Dhama Coupling and lines of code as independent variables are organized in Table 1 (Cuauhtemoc et al. 2005):

| Pair        | \(r\)  |
|-------------|--------|
| MC_DC       | -0.3860 |
| MC_LOC      | 0.7653 |
| DC_LOC      | -0.4346 |

Table 1. Correlation between variables

| Variable Name | Range | MF     | Parameters a | b | c |
|---------------|-------|--------|--------------|---|---|
| McCabe        | 1-7   | Low 1  | 2 3          |
|               |       | Average 2 | 4 5          |
|               |       | High 4 | 6 7          |
| Dhama         | 0-0.4 | Low 0.24 | 0.31 0.40 |
|               |       | Average 0.05 | 0.17 0.33 |
|               |       | High 0.00 | 0.12 0.21 |
| Lines of Code | 1-40  | Low 3  | 8 15         |
|               |       | Average 10 | 16 25        |
|               |       | High 21 | 28 40        |

| Variable Name | Range | MF     | Parameters a | b | c |
|---------------|-------|--------|--------------|---|---|
| Development Time (min) | 1-27 | Low 6.6 | 9.0 11.8 | |
|               |       | Average 8.1 | 12.8 18.6 |
|               |       | High 14.0 | 20.0 27.0 |

Table 2. Membership Function Parameters

So fuzzy rules were formulated as bellow:

1. If Complexity is low and Size(LOC) is small then DT is low
2. If Complexity is average and Size(LOC) is medium then DT is average
3. If Complexity is high and Size(LOC) is big then DT is high
4. If Coupling is low then DT is low
5. If Coupling is average then DT is average
6. If Coupling is high then DT is high

Fig. 1. Membership functions for input & output
For triangular membership function, three parameters (a, b, c) are defined. In Table 2 (Cuauhtemoc et al. 2005), input and output membership function is shown for dependent and independent variables. Their scalar parameters (a, b, c) are defined as follows:

\[ MF(x)=0 \text{ if } x < a \]
\[ MF(x)=1 \text{ if } x = b \]
\[ MF(x)=0 \text{ if } x > c \]

The membership functions corresponding to Table 2 are shown in Fig. 1(a), 1(b), 1(c), and 1(d).

Consequently, by using fuzzy rules and their memberships, DT is depicted in Table 4.

4. Multiple Regression Model with Fuzzy Concepts

Fuzzy concepts help us to find the deviation of each data from fitness equation, so we define a normal distribution membership function as follow:

\[ U_i = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}} \]  

(5)

Where \( \mu \) is average of sample points and \( \sigma \) is square root of variance math.

If we add fuzzy domain to Regression method, the effect of discrete data points on the fitness result will be reduced and the effect of concentrated data points on the fitness result will be enhanced.

For each data in Table 3, we obtain the membership function that is shown in column 7. A group of equations to obtain the fuzzy parameters are given as (Gu et al. 2006):

\[ s_{j1}b_1 + s_{j2}b_2 + \ldots + s_{jk}b_k = s_{ij} \]
\[ s_{j1}b_1 + s_{j2}b_2 + \ldots + s_{jk}b_k = s_{ij} \]
\[ \vdots \]
\[ s_{j1}b_1 + s_{j2}b_2 + \ldots + s_{jk}b_k = s_{ij} \]  

(6)

For each Development Time (DT) of mentioned projects, and here we have 41 projects for considering.

Then \( b_0 \) is calculated by:

\[ b_0 = \sum_{i=1}^{N} \sum_{j=1}^{k} (y_i - \bar{y}) (x_{ij} - \bar{x}) / \sum_{i=1}^{N} \sum_{j=1}^{k} (x_{ij} - \bar{x})^2 \]  

(7)

By solving these equations, final equation is expressed as:

\[ DT = 17.3352512 + 1.709789562 * MC - 29.5519386 * DC - 0.0370081952 * LOC \]  

(8)

The result of this method is presented in Table 4.

5. Evaluating Techniques

A common criterion, which is calculated for each observation, is MRE and it is defined as follows:

\[ MRE = \frac{|Actual \ Effort_i - Predicted \ Effort_i|}{Actual \ Effort_i} \]  

With aggregation of MRE on all data, MMRE (Mean Magnitude of Relative Error) is achieved as follows (Burgess & Lefley, 2001):

\[ MMRE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{E_i - \bar{E}}{E_i} \right| \]  

(10)

A complementary criterion that is used here is Pred(20).

In general, Pred(l)=k/N where k is the number of observations where MRE is less than or equal to l (Cuauhtemoc et al. 2006). So Pred(20) gives the
percentage of projects which were predicated with a MRE less or equal than 0.20. In general, the accuracy of an estimation technique is proportional to the Pred(20) and inversely proportional to the MMRE (Xia et al. 2005).

6. Experimental Results

Multiple Regression, fuzzy rules system and fuzzy multiple regression are applied to the same data subset. The MMRE & Pred(20) are shown in Table 4. Results are indicated that fuzzy multiple regression model is better than linear regression equations and fuzzy models in both evaluation criterion (Pred(20) & MMRE). Comparison between actual development time, Multiple Regression Model, Fuzzy Logic and Fuzzy Multiple Regression Model is shown in Fig. 2. This figure is showed that fuzzy multiple regression output is close to actual development time in compare to the other models.

7. Conclusions and Future Research

The goal of this paper is to investigate the models for estimating software project. These techniques have been compared in terms of accuracy. Research demonstrates that fuzzy multiple regression models are better than linear regression equations and fuzzy models. An ongoing research is related to apply neural network models using Bayesian Regularization training algorithm to data subset, because is more stable than fuzzy models that have membership functions whose derivatives have discontinuities at some points.

8. References

Moataz A. Ahmed, Moshood Omolade. Saliu, J. AlGhamdi, “Adaptive fuzzy logic-based framework for software development effort prediction”, Elsevier Science, Information and software technology, pp 31-48, 2005.
V.Xia, D.Ho, L.F.Capretz, “Calibrating Function Points Using Neuro-Fuzzy Technique”, 2005.
S.G.MacDonell, “Software source code sizing using fuzzy logic modelling”, Elsevier Science, Information and software technology, pp 389-404, 2003.
L.M.Cuaughtemoc, J.Lebeouf, M.C.Yanez, T.Agustin Gutierrez, “Software Development Effort Estimation Using Fuzzy Logic: A Case Study”, IEEE, Proceedings of the Sixth Mexican International Conference on Computer Science (ENC’05), 2005.
X.Gu, G.Song, L.Xiao, “Design of a Fuzzy Decision-making Model and Its Application to Software Functional Size Measurement”, IEEE, International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06), 2006.
C.J.Burgess, M.Lefley, “Can genetic programming improve software effort estimation? A comparative evaluation”, Elsevier Science, Information and software technology, pp 863-873, 2001.
L.M.Cuaughtemoc, Y.M Cornelio, G.T Agustin, “A Fuzzy Logic Model Based Upon Reused and New & Changed Code for Software Development Effort Estimation at Personal Level”, IEEE, Proceedings of the 15th International Conference on Computing (CIC’06), 2006.
H.B. Kuan Tan, Y.Zhao, H. Zhang, “Estimating LOC for Information Systems from their Conceptual Data Models”, ACM, 2006.
O.Lima, P. P. Muniz Farias, A.D.Belchior,” A Fuzzy Model for Function Point Analysis to Development and Enhancement Project Assessments”, CLEI Electronic Journal, Vol.5, No.2, 1999.