ESTIMATING PROJECT DEVELOPMENT EFFORT USING CLUSTERED REGRESSION APPROACH

Geeta Nagpal[1], Moin Uddin[2] and Arvinder Kaur[3]

[1]Dept. of Computer Science and Engg,
National Institute of Technology, Jalandhar
sikkag@gmail.com
[2]Pro Vice Chancellor, Delhi Technological University, Delhi
prof_moin@yahoo.com
[3]University School of Information Technology, GGSIPU, Delhi
arvinderkaurtakkar@yahoo.com

ABSTRACT

Due to the intangible nature of "software", accurate and reliable software effort estimation is a challenge in the software Industry. It is unlikely to expect very accurate estimates of software development effort because of the inherent uncertainty in software development projects and the complex and dynamic interaction of factors that impact software development. Heterogeneity exists in the software engineering datasets because data is made available from diverse sources. This can be reduced by defining certain relationship between the data values by classifying them into different clusters. This study focuses on how the combination of clustering and regression techniques can reduce the potential problems in effectiveness of predictive efficiency due to heterogeneity of the data. Using a clustered approach creates the subsets of data having a degree of homogeneity that enhances prediction accuracy. It was also observed in this study that ridge regression performs better than other regression techniques used in the analysis.

KEYWORDS

Software estimation, Clustering, Grey relational analysis, Feature weighted Grey relational based clustering

1. INTRODUCTION

Development of software is a creative process where each person's efficiency is different. It is initially difficult to plan and estimate as most software projects have deficient information and vague associations amongst effort drivers and the required effort. Software developers and researchers are using different techniques and are more concerned about accurately predicting the effort of the software product being developed.

Most commonly used algorithmic models for software cost estimation include Boehm's COCOMO [1], Albrecht's Function Point Analysis [2], and Putnam's SLIM [3]. These models require some cost drivers to estimate the effort of the project. In recent years, a number of soft computing and computationally intelligent techniques have been proposed in order to handle the unpredictability and inherent uncertainty contributed by cost drivers' judgment and environment

Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA - 2013
pp. 493–507, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3654

complexity of the projects. These techniques include artificial neural network, genetic algorithm, support vector regression, genetic programming, neuro-fuzzy inference system and case base reasoning. They use the historical datasets of completed projects as training data and predict the values for new project's effort based on previous training. Though, a significant improvement has been achieved using soft computing techniques in software cost estimation, yet there exist some limitations due to the heterogeneity in the datasets.

Soft computing techniques estimate accurately if there is some relationship between the tuples of the dataset. Due to this heterogeneity that exists amongst software projects, these techniques are not able to estimate optimally. This heterogeneity of data can be reduced by clustering the data into some similar groups. The goal of clustering is to create the groups of data that have similar characteristics. The clustering divides the data set X into k disjoint subsets that have some dissimilarity between them.

A clustered regression approach has been proposed in this study in order to generate more efficient estimation sub models. A feature weighted grey relational based clustering method has been integrated with regression techniques. The feature weighted grey relational clustering algorithm uses grey relational analysis for weighting features and also for clustering. The results obtained showed that clustering could decrease the effect of irrelevant projects on accuracy of estimations. Cluster specific regression models for the four publicly available data sets are generated. Empirical results have shown that regression when applied on clustered data provides some outstanding results, thus indicating that the methodology has great potential and can be used for software effort estimation. The results are subjected to statistical testing using Wilcoxon signed rank test and Mann_Whitney U Test.

The rest of the paper is organized as follows. Section 2 reviews some related works on clustering algorithm and GRG as a similarity measure for feature weighting. Section 3, introduces the modeling techniques. Further, in Section 4 we present the proposed methodology. Section 5, gives description of the data sets used in the study and the experimental results that demonstrate the use of proposed clustered regression approach in software effort estimation. In the end the conclusion is made in Section 6.

2. REVIEW OF LITERATURE

A number of data clustering techniques have been developed to find the optimal subsets of data from the existing datasets [4],[5],[6]. The main aim of clustering is to partition an unlabeled data set into subsets according to some similarity measure, called unsupervised classification. Clustering algorithms can be categorized into two main families: input clustering and input-output clustering [7]. In input clustering algorithms all the attributes are considered as independent. Hard c-means [8] and fuzzy c-means [9] algorithms fall into this category. In the case of input-output clustering each multi-attribute data point is considered as a vector of independent attribute with some corresponding dependent value. Let $S = \{(x_1,y_1), (x_2,y_2) ,\ldots (x_n, y_n)\}$ be a set of unlabelled input-output data pairs. Each independent input vector $x_i = [x_{1i}, x_{2i},\ldots x_{ki}]$ has a corresponding dependent value y_i. Research work has been done to motivate this category of classification [10],[11].

Kung and Su [12] developed an effective approach to establish affine Takagi-Sugeno (T-S) fuzzy model for a nonlinear system from its input – output data. Chunheng, Cui and Wang [13] proposed FCM-SLNMM clustering algorithm, consisting of two stages. The FCM algorithm was applied in the first stage and supervised learning normal mixture model was applied in the second stage. The clustering results of the first stage were used as training data. The experiments on the real world data from the UCI repository showed that the supervised learning normal mixture

model can improve the performance of the FCM algorithm sharply. Lin and Tsai [14] proposed a hierarchical grey clustering approach in which the similarity measure was a globalized modified grey relational grade instead of traditional distances. Chang and Yeh [15] generalized the concept of grey relational analysis in order to develop a technique for analyzing the similarity between given patterns. They also proposed a clustering algorithm to find cluster centers of a given dataset.

In this study, GRA a technique of GST which utilizes the concept of absolute point-to-point distance between features [16],[29] has been applied. GST a recently developed system engineering theory, was first established by Deng [18],[19],[20]. It draws out valuable information by generating and developing the partially known information. So far, GST has been applied in different areas of image processing [23], mobile communication [24], machine vision inspection [25], decision making [26], stock price prediction [27] and system control [28]. The success of GST motivated us to investigate its application in software effort estimation.

3. MODELING TECHNIQUES

The data available for software cost estimation is inherently non linear and hence accurate estimation of effort is difficult. Efficient estimation can be achieved if this non linearity that exists can be treated by tracing relationships among data values. In this study, we try to reduce the heterogeneity by applying feature weighted grey relational clustering methodology.

3.1 Regression

As discussed, a large number of techniques have been applied to the field of software effort estimation. The aim of this study is to assess which regression techniques perform best to estimate software effort. The following techniques are considered.

3.1.1 Ordinary Least Square Regression

It is the most popular and widely applied technique to build software cost estimation models. According to principle of least squares the `best fitting' line is the line which minimizes the deviations of the observed data away from the line. The regression parameters for the least square line, are estimates of the unknown regression parameters in the model. This is referred to as multiple linear regression and is given by:

$$y_i = \beta_0 + \beta_1 x_{i,1} + ...\beta_k x_{i,k} + \varepsilon_i \quad (1)$$

where, y_i is a dependent variable where as $x_1, x_2, \ldots\ldots x_k$ are k independent variables. β_o is the y intercept, β_1, β_2 are the slope of y, ε_i is the error term.

3.1.2 Ridge Regression (RR)

RR is an alternative regression technique that tries to address the potential problems with OLS that arise due to highly correlated attributes. In regression, the objective is to “explain” the variation in one or more “response variables”, by associating this variation with proportional variation in one or more “explanatory variables”, but the problem arises when the explanatory variables vary in similar ways reducing , their collective power of explanation. The phenomenon is known as *near collinearity*. As the different variables are correlated the covariance matrix XX will be nearly singular and as a result the estimates will be unstable. A small variation in error will have large impact on $\hat{\beta}$. Ridge regression reduces the sensitivity by adding a number δ to

the elements on the diagonal of the matrix to be inverted. δ is called the ridge parameter and it yields the following estimator of β.

$$\hat{\beta}_{\delta} = (X'X + \delta I_n)^{-1}(X'e) \tag{2}$$

where, I_n represents the identity matrix of rank n.

3.1.3 Forward Stepwise Regression

The purpose of stepwise regression is to generate regression model in which the detection of most predictive variables is carried out. It is carried out by a series of F tests. The method evaluates the independent variables at each step, adding or deleting them from the model based on user-specified criteria. In the first step, each of the independent variables are evaluated individually and the variable that has the largest F value greater than or equal to the F to enter value, is entered into the regression equation. In the subsequent steps, when a variable is added to the model based on their F value, the method also examines variables included in the model based on F to remove criteria, and if any variables are found they are removed.

3.1.4 Backward Stepwise Regression

The backward stepwise elimination procedure is basically a series of tests for significance of independent variables. The process starts with the maximum model, it eliminates the variable with the highest *p* -value for the test of significance of the variable, conditioned on the *p* -value being bigger than some pre-determined level (say, 0.05). In the next step, it fits the reduced model after having removed the variable from the maximum model, and also removes from the reduced model the variable with the highest *p*-value for the test of significance of that variable (if $p>=0.05$) and so on. The process ends when no more variables can be removed from the model at significance level 5%.

3.1.5 Multiple Adaptive Regression Splines(MARS)

MAR splines focuses on the development and deployment of accurate and easy-to-understand regression models. The MAR splines model is designed to predict continuous numeric and high quality probability models. MAR spline model is a regression model which automatically generates non-linearities and interactions between variables and is thus a promising technique to be used for software effort estimation[21]. MAR splines fits the data to the following equation.

$$e_i = b_0 + \sum_{k=1}^{K} b_k \Pi_{i=1}^{L} h_i(x_i(j)) \tag{3}$$

In this b_o and b_k are the intercept and slope. Parameters $h_i(x_i(j))$ are the hinge functions. They take the form *max (0, x_i (j)-b)* where, *b* is the knot. MAR splines is a multiple piece wise linear regression by adding multiple hinge functions.

3.2 Grey Relational Analysis

GRA is comparatively a novel technique in software estimations. It is used for analyzing the relationships that exists between two series. The magnetism of GRA to software effort estimation shoots from its flexibility to model complex nonlinear relationship between effort and cost drivers [16].

Grey Relational Grade by Deng's Method [18],[19],[20]

GRA is used to quantify all the influences of various factors and the relationship among data series that is a collection of measurements [16].The three main steps involved in the process are:

Data Processing: The first step is the standardization of the various attributes. Every attribute has the same amount of influence as the data is made dimensionless by using various techniques like upper bound effectiveness, lower bound effectiveness or moderate effectiveness.

Upper-bound effectiveness (i.e., larger-the-better) is given by:

$$x_i^*(k) = \frac{x_i(k) - \min_i x_i(k)}{\max_i x_i(k) - \min_i x_i(k)} \tag{4}$$

where i=1,2,…,m and k=1,2,…,n.

Difference Series: GRA uses the grey relational coefficient to describe the trend relationship between an objective series and a reference series at a given point in a system.

$$\gamma(x_0(k), x_i(k)) = \frac{\Delta_{\min} + \zeta\Delta_{\max}}{\Delta_{o,i}(k) + \zeta\Delta_{\max}} \tag{5}$$

where;

$\Delta_{0,i}(k) = |x_0(k) - x_i(k)|$ is the difference of the absolute value between $x_0(k)$ and $x_i(k)$;

$\Delta_{min} = min_j min_k |x_0(k) - x_j(k)|$ is the smallest value of $\Delta_{0,j}$ $\forall j \in \{1, 2, \ldots, n\}$;

$\Delta_{max} = max_j max_k |x_0(k) - x_j(k)|$ is the largest value of $\Delta_{0,j}$ $\forall j \in \{1, 2, \ldots, n\}$; and

ζ is the distinguishing coefficient, $\zeta \in (0, 1]$.

The ζ value will change the magnitude of $\gamma(x_0(k), x_i(k))$. In this study the value of ζ has been taken as 0.5 [17].

Grey Relational Grade: GRG is used to find overall similarity degree between reference tuple x_o and comparative tuple x_i. When the value of GRG approaches 1, the two tuples are "more closely similar". When GRG approaches a value 0, the two tuples are "more dissimilar". The GRG $\Gamma(x_0, x_i)$ between an objective series x_i and the reference series x_0 was defined by Deng as follows:

$$\Gamma(x_0, x_i) = \frac{1}{n}\sum_{k=1}^{n} \gamma(x_0(k), x_i(k)) \tag{6}$$

4. PROPOSED METHODOLOGIES

4.1 Clustered regression using grey relational analysis

In this methodology, in order to reduce the heterogeneity that exists in the datasets, the initial focus is to partition datasets into subsets according to some similarity measure, called unsupervised classification. The proposed clustering algorithm uses grey relational analysis for feature selection as well as for clustering. In this the maximum mean grey relational grade

between data points acts as an objective function instead of the minimum distance used by *k*-means. The structural flow chart is shown in figure 1.

The three main steps involved in the algorithm are:
Step 1: Using Grey relational analysis for finding feature weights.
Step 2: Using Grey relational analysis for clustering the datasets based on these feature-weights.
Step 3: Applying regression techniques on the clustered datasets.
Step 4: Effort Prediction by Regression Techniques

The detailed algorithm is described as follows:

Using Grey Relational Analysis for finding feature weights.

Feature Selection by GRA [16]

a. *Construction of data:* The columns in each cluster dataset are treated as series . The effort series $x_e = \{e_1, e_2, e_3 ..... e_n\}$ is taken as the reference series and the attribute columns are regarded as objective series.
b. *Normalization:* Each data series is normalized according to equation 4, so that they have same degree of influence on the dependent variable "effort".
c. *Generation of Grey Relational Grade:* Grey relational grade(GRG) is calculated for each series wrt reference series according to equation 6.

The GRG's are generated, normalized and used as the corresponding features weight w_k.

Using Grey Relational Analysis for clustering the datasets based on these feature-weights.

After finding the weights of the features from the first step it applies the clustered approach based on grey relational analysis. The detailed algorithm is described as follows:

a. Weight of each feature is generated as described earlier.
b. Normalize the data with larger the better as per equation 4,
c. Calculate distance between data points based on weighted GRG
   - Consider the i^{th} data point as reference series x_o
   - All the other data points as the objective series, $\{ x_1, x_2, x_3, \ldots\ldots x_{n-1} \}$
   - Calculate the grey relational coefficient with ζ=0.5 [17].Calculate weighted GRG of the reference series and feature weight calculated as in step 1 for all objective series.
d. Randomly select the number of desired of clusters center c_k.
e. GRG distance from data points to cluster centers is used as a basis to select cluster members, for which it has maximum GRG value.
f. Update the cluster centers by selecting centers based on the maximum mean GRG, then repeat step *e*.
g. Repeat steps *e* and *f*, until there is no change in the cluster head updating or the difference between the mean is less than some predefined threshold value.

5. EXPERIMENTAL RESULTS

5.1 Dataset

In order to evaluate the models based upon the proposed methodology, four well established datasets from the Promise repository [22] have been used for validating our model. These datasets are Desharnais, Finnish, Albrecht and Maxwell. The descriptive statistics of the datasets are shown in Table 1 given below. All the datasets have a varied range of effort values. These

datasets have been treated individually as they have distinct features. Also the clusters from each dataset have been treated separately. The prediction accuracy for all the models with and without clustering are then compared. In order to measure the accuracy of the software estimation, we have used three most popularly used evaluation criteria in software engineering *i.e MMRE*, *MdMRE* and *Pred(n).*

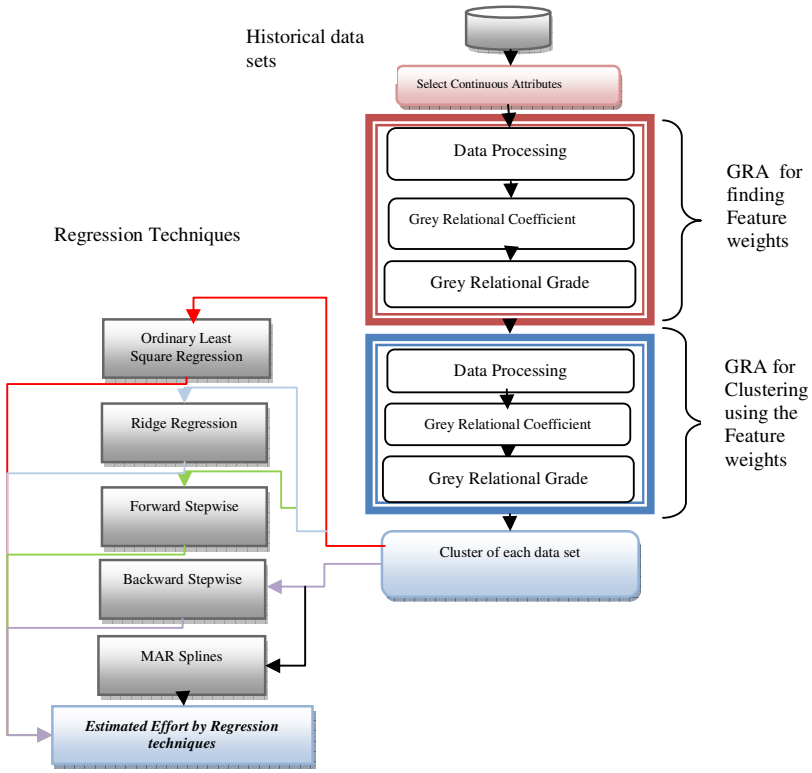

Figure1. Structural flowchart for feature weighted Grey Relational Clustering

Table1. Descriptive Statistics of the data sets

	Cases	Features	Effort Mean	Minimum (effort value)	Maximum (effort value)	Effort Std. Dev.
Albrecht	24	8	21.875000	0.50	105.20	28.417895
Desharnais	77	11	4833.9090	546	23940	4188.1851
Finnish	38	8	7678.2894	460	26670	7135.2799
Maxwell	62	23	8223	583	63694	10499.903

5.1.1 Comparison over Desharnais data set

The results obtained suggest that applying regression technique on clustered data produces more accurate estimation models than applying regression on the entire datasets. This is evident from the results obtained shown in Table 2. The *Pred(25)* accuracy has improved from 35.06 % to 50 % using OLS regression whereas, the *MMRE* and *MdMRE* has fallen from 0.5 to 0.32 and from 0.31 to 0.25 respectively. Similar observations can be notified from the table below for all other regression models also. Best results have been obtained on using the proposed feature weighted grey relational clustering.

Table 2. Prediction accuracy results (Desharnais data set)

	OLS	Ridge Regression	Forward Stepwise	Backward Stepwise	MAR Splines
Desharnais					
MMRE	0.5	0.47	0.5	0.5	0.51
MdMRE	0.31	0.3	0.31	0.31	0.32
Pred(25)	35.06	41.56	37.66	37.66	35.06
Desharnais(Cluster_1) using Grey Relational Clustering					
MMRE	0.32	0.35	0.33	0.39	0.39
MdMRE	0.25	0.23	0.24	0.21	0.20
Pred(25)	50	55.56	50	52.78	58.33

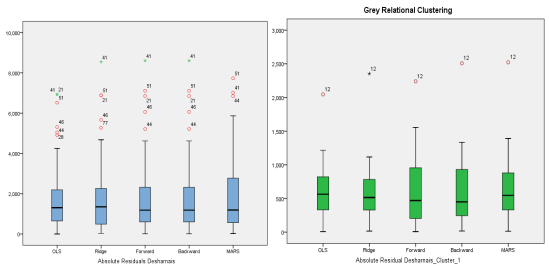

Figure 2. Boxplot of Absolute Residuals for Desharnais

The Boxplot of absolute residuals provide good indication of the distribution of residuals and can help better understand the mean magnitude of relative error and *Pred (25)*.The results obtained were subjected to statistical tests using Wilcoxon Signed rank test and Mann Whitney U test.
The box plots of absolute residuals shown in figure 2 suggest that:

The medians for all regression techniques applied on Desharnais _Cluster are more close to zero, as it is clear from the values on the Y-axis, indicating that the residuals were closer to the minimum value. The Outliers are few and less extreme in cases of Desharnais_Clusters as compared to Desharnais data set. As the *p*-value in all the cases shown in Table 3 is greater 0.05 where in we conclude that the residuals obtained in all approaches are not significantly different from the test value zero. As a result, the proposed methods can be used for software effort estimation. The statistical tests were performed using SPSS 19 for windows.

Table 3. Wilcoxon signed rank test Test

	Desharnais		**Desharnais_Cluster_1 using Grey Relational**	
	Z	Asymp. Sig. (2-tailed)	Z	Asymp. Sig. (2-tailed)
OLS-Actual	-.419a	.675	-.236a	.814
Ridge – Actual	-.551a	.582	-.299a	.765
Forward - Actual	-.449a	.653	-.189a	.850
Backward- Actual	-.449a	.653	-.314a	.753
MARS– Actual	-.566a	.571	-.236a	.814

The results of Mann_Whitney U Test are provided in Table 4. Predictions obtained using the clustered approach presented statistically significant estimations.

Table 4. Results Mann-Whitney U Test

	Desharnais vs. Desharnais_Cluster_1 using Grey Relational Z
OLS Regression	-4.018
Ridge Regression	-4.079
Forward Stepwise	-4.252
Backward Stepwise	-4.227
MAR Splines	-4.240

5.1.2 Comparison over Finnish data set:

For the Finnish dataset, some significant results (as shown in Table 5.) were obtained on the clustered data. The *Pred(25)*accuracy improved from 36.84 % to 100 % using OLS regression whereas, the *MMRE* and *MdMRE* has fallen from 0.75 to 0.02 and from 0.36 to 0.02 respectively. Similar observations can be notified from the table below for all other regression models also.
The boxplot of absolute residuals for Finnish dataset and Finnish_cluster is shown in Figure 3. They suggest that:

The medians for all regression techniques applied on Finnish _Cluster are very close to zero, as it is clear from the values on the Y-axis, indicating that the estimates were closer to the minimum value. Outliers are less extreme in case of Finnish_Cluster. One sample Wilcoxon signed rank test has been applied in order to investigate the significance of the results by setting level of confidence to 0.05. From the results obtained as shown in Table 6, we can conclude that no significant difference exists between the residual median and hypothetical median.

Table 5. Prediction accuracy results (Finnish data set)

	OLS	**Ridge Regression**	**Forward Stepwise**	**Backward Stepwise**	**MAR Splines**
			Finnish		
MMRE	0.75	0.71	1.01	0.76	0.08
MdMRE	0.36	0.32	0.43	0.42	0.07
Pred(25)	36.84	36.84	36.84	36.84	97.37
			Finnish(Cluster_1) using Grey Relational Clustering		
MMRE	0.02	0.025	0.23	0.022	0.022
MdMRE	0.02	0.024	0.022	0.023	0.023
Pred(25)	100	100	100	100	100

Unsurprisingly, predictions based on clustered regression model presented statistically significant accurate estimations, measured using absolute residuals, confirmed by the results of boxplot of absolute residuals and also verified using Mann-Whitney U test (Table 7.)

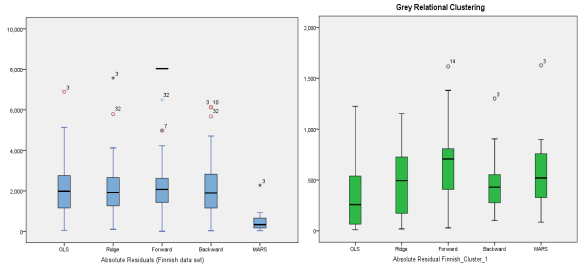

Figure 3. Boxplot of Absolute Residuals for Finnish

Table 6. Wilcoxon signed rank test

	Finnish		Finnish_Cluster_1 using Grey Relational	
	Z	AsympSig (2-tailed)	Z	Asymp.Sig. (2tailed)
OLS- Actual	-.268a	.788	-.408a	.683
Ridge – Actual	-.355a	.722	-.220b	.826
Forward- Actual	-.268a	.788	-.157b	.875
Backward- Actual	-.152a	.879	-.345a	.730
MARS– Actual	-.558a	.577	-.282a	.778

Table 7. Results Mann-Whitney U Test

	Finnish vs. Finnish_Cluster_1 using Grey Relational Z
OLS Regression	-2.022
Ridge Regression	-2.104
Forward Stepwise	-2.228
Backward Stepwise	-2.022
MAR Splines	-1.939

5.1.3 Comparison over Albrecht data set:

The results obtained using the proposed clustered regression approach produced more accurate models. This is evident from the *Pred(25)* accuracy that improved from 37.5 % to 85.71 % using OLS whereas, the *MMRE* and *MdMRE* has fallen from 0.9 to 0.09 and from 0.43 to 0.05 respectively. Similar observations can be notified for all other regression models also (Table 8).

Table 8. Prediction accuracy results (Albrecht data set)

	OLS	Ridge Regression	Forward Stepwise	Backward Stepwise	MAR Splines
			Albrecht		
MMRE	0.9	0.91	0.86	1	1.23
MdMRE	0.43	0.52	0.5	0.49	0.6
Pred(25)	37.5	37.5	41.67	37.5	29.17
	Albrecht(Cluster_1) using Grey Relational clustering				
MMRE	0.092	0.21	0.19	0.08	0.225
MdMRE	0.05	0.24	0.22	0.025	0.175
Pred(25)	85.71	50	57.14	85.71	57.14

The box plots of absolute residuals suggest that:

The medians for all regression techniques applied on Albrecht_Clusters are very close to zero, as it is clear from the values on the Y-axis, indicating that the residuals were closer to the minimum value. Outliers are less extreme in case of Finnish_Cluster.

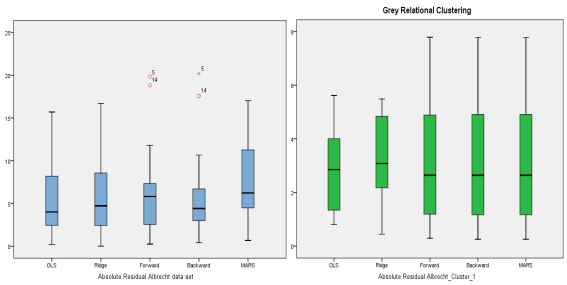

Figure 4. Boxplot of Absolute Residuals for Albrecht

The results of Wilcoxon signed rank conclude that no significant difference exists between the residual median and hypothetical median, thus indicating good predictions (Table 9.)

Table 9. Wilcoxon signed rank test

	Albrecht		**Albrecht_Cluster_1 using Grey Relational**	
	Z	Asymp.Sig. (2-tailed)	Z	Asymp.Sig (2-tailed)
OLS- Actual	-.029a	.977	-.031a	.975
Ridge – Actual	-.057a	.954	-.031a	.975
Forward - Actual	-.057a	.954	-.031a	.975
Backward Actual	-.086b	.932	-.031a	.975
MARS– Actual	-.029b	.977	-.031a	.975

Table 10. Results Mann-Whitney U Test

	Albrecht vs. Albrecht_Cluster_1 using Grey Relational Z
OLS Regression	-1.014
Ridge Regression	-1.014
Forward Stepwise	-.943
Backward Stepwise	-1.155
MAR Splines	-1.108

The results obtained using Mann-Whitney U test shown in Table 10, however didn't prove significant difference between the proposed approaches. This is because of the small size of the dataset. The data set comprises of 24 projects. It was divided into two clusters one with 20 projects and other with 4 projects. Clustered regression approach was applied on 20 projects.

5.1.4 Comparison over Maxwell Dataset

The results obtained in Table 11 using the proposed clustered regression approach produced more accurate models for Maxwell dataset also. This is evident from the Pred(25) accuracy that improved from 38.71 % to 51.51 % using OLS regression whereas, the *MMRE* and *MdMRE* has fallen from 0.59 to 0. 51 and from 0.38 to 0.24 respectively. For Ridge regression also, the *Pred(25)* accuracy has increased from 43.55% to 60.60% which is significant improvement, the *MMRE* and *MdMRE* has gone low from 0.54 to 0.33 and 0.3 to 0.18 respectively.

Table 11. Prediction accuracy results(Maxwell data set)

	OLS	Ridge Regression	Forward Stepwise	Backward Stepwise	MAR Splines
			Maxwell		
MMRE	0.59	0.54	0.53	0.59	0.7
MdMRE	0.38	0.3	0.32	0.33	0.46
Pred(25)	38.71	43.55	38.71	37.1	32.26
			Maxwell(Cluster_1) using Grey Relational Clustering		
MMRE	0.51	0.33	0.46	0.60	0.75
MdMRE	0.24	0.18	0.25	0.25	0.52
Pred(25)	51.51	60.60	48.48	48.48	24.24

The box plots of absolute residuals suggest that:

The medians for all regression techniques applied on Maxwell _Cluster are more close to zero, as it is clear from the values on the Y-axis, indicating that the estimates were closer to the minimum value. The medians are more skewed to the minimum value indicating that the predictions are good. Outliers are few and less extreme in case of Maxwell_Cluster as compared to entire dataset.

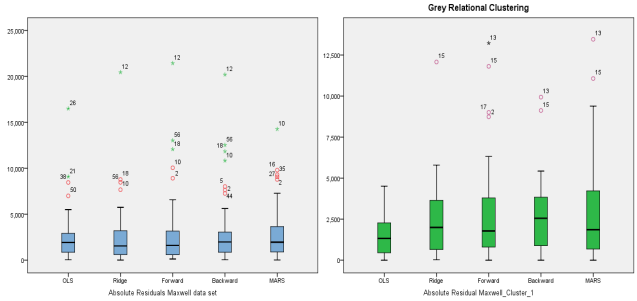

Figure 5. Boxplot of Absolute Residuals for Maxwell

Table 12. Wilcoxon signed rank test

	Maxwell		Maxwell_Cluster_1 using Grey Relational	
	Z	Asymp. Sig (2-tailed)	Z	Asymp. Sig (2-tailed)
OLS-Actual	-249a	.803	-.068a	.946
Ridge – Actual	-.508a	.611	-.616b	.538
Forward – Actual	-.691a	.490	-.068b	.946
Backward-Actual	-.831a	.406	-.023b	.982
MARS – Actual	-.109a	.913	-.205a	.838

Results of Wilcoxon signed rank test suggest that no significant difference exists between the residual median and hypothetical median. The results of Wilcoxon signed rank test are given in Table 12.

Concerning the statistical test based on Mann-Whitney U(Table 13), we found no significant difference between clustered regression approach and regression approach.

Table 13. Results Mann-Whitney U Test

	Maxwell vs. Maxwell_Cluster_1 using Grey Relational Z
OLS Regression	-1.431
Ridge Regression	-1.448
Forward Stepwise	-1.482
Backward Stepwise	-1.312
MAR Splines	-1.806

6. CONCLUSION

This work resolves the heterogeneity problems that exist in the datasets. In order to confirm the effectiveness of proposed work, four different data sets have been used for software estimation. Simulation results obtained provide a comparison of clustered regression approach over only regression. The results confirm that the proposed feature weighted grey relational clustering algorithm performed appreciably for software effort estimation. The statistical test based on Mann-Whitney U, further confirmed that statistical significant difference exists between the proposed clustered-regression models and regression models.

Further, this work can be extended by using clustered approach with different soft computing techniques with different similarity measures for feature selection. GRA can also be analyzed for feature subset selection Also, for enhanced efficiency in software estimation the techniques should be applied on large data sets with different clustering algorithms.

REFERENCES

[1] Boehm, B (1981) Software Engineering Economics Englewood Cliffs, NJ, Prentice Hall.
[2] Albrecht, A.J. & Gaffney, J.R. (1983) “Software measurement, source lines of code, and development effort prediction: a software science validation”, IEEE Transactions on Software Engineering, Vol. 9, No. 6, pp 639-648.
[3] Putnam, Lawrence H. (1978) "A General Empirical Solution to the Macro Software Sizing and Estimating Problem", IEEE Transactions on Software Engineering, Vol. SE-4, No. 4, pp 345-361.
[4] El-Zaghmouri, B. M. & Abu-Zanona, M. A. (2012) “Fuzzy C-Mean Clustering Algorithm Modification and Adaption for Application”, World of Computer Science and Information Technology Journal, ISSN: 2221-0741, Vol.2, No.1, pp 42-45.
[5] Lin, C. T. & Tsai, H. Y. (2005) “Hierarchical Clustering Analysis Based on Grey Relation grade”, Information and Management Sciences, Vol. 16, No. 1, pp 95-105.
[6] Wong, C.C. & Chen, C.C. (1998) “Data clustering by grey relational analysis”, J. Grey Syst, Vol. 10, No. 3, pp 281-288.
[7] Hu, Y.C., Chen, R. S., Hsu, Y. T., & Tzebg, G. H. (2002) “Grey self-organizing feature maps”, Neuro computing, Vol. 48, No.1–4, pp 863-877.
[8] Duda, R.O., & Hart, P.E. (1973) Pattern classification and scene analysis, John Wiley &Sons, Inc., New York.
[9] Bezdek, J. C., Ehrlich, R. & Full, W. (1984) “FCM: The Fuzzy c- Means Clustering Algorithm”, Computers & Geoscience, Vol. 10, No. 2-3, pp 191-203.
[10] Runkler, T.A. & Bezdek, J.C. (1999) “Alternating cluster estimation: a new tool for clustering and function approximation”, IEEE Trans. Fuzzy Syst., Vol. 7, No. 4, pp 377-393.
[11] Pedrycz, W.(1996) “Conditional fuzzy c-means”, Pattern Recogn. Lett.,Vol. 17, No. 6, pp 625-632.
[12] Kung C. C & Su J. Y. (2007) “Affine Takagi-Sugeno fuzzy modeling algorithm by Fuzzy c-regression models clustering with a novel cluster validity criterion”, IET Control Theory Appl., pp. 1255 – 1265.

[13] Wang, W., Wang, C., Cui, X. & Wang, A. (2008) "A clustering algorithm combine the FCM algorithm with supervised learning normal mixture model", ICPR 2008, pp 1-4.
[14] Lin, C. T. & Tsai, H. Y. (2005) "Hierarchical Clustering Analysis Based on Grey Relation grade", Information and Management Sciences, Vol. 16, No. 1, pp 95-105.
[15] Chang, K. C. & Yeh, M. F. (2005) "Grey Relational Based Analysis approach for data clustering", IEEE Proc.-Vis. Image Signal Process, Vol.152, No.2.
[16] Song, Q., Shepperd M., Mair C. (2005) "Using Grey Relational Analysis to Predict Software Effort with Small Data Sets". Proceedings of the 11th International Symposium on Software Metrics (METRICS'05), pp 35-45.
[17] Azzeh, M., Neagu, D. & Cowling, P. I., (2010) "Fuzzy grey relational analysis for software effort estimation", Journal of Empirical software Engineering, Vol.15, No.1, **[doi:**10.1007/s10664-009-9113-0]
[18] Deng, J. L. (1982) "Control problems of grey system", System and Control Letters, Vol. 1, pp 288-94.
[19] Deng, J. (1989) "Introduction to Grey System theory", The Journal of Grey System, Vol.1, No.1, pp 1-24.
[20] Deng, J. (1989) "Grey information space", The Journal of Grey System, Vol.1, No.1, pp 103-117.
[21] MATLAB® Documentation, http://www.mathworks.com/help/techdoc/
[22] PROMISE Repository of empirical software engineering data http://promisedata.org/ repository.
[23] Jou, J. M , Chen, P. Y & Sun, J. M. (1999) "The gray prediction search algorithm for block motion estimation", IEEE Transactions on Circuits and Systems for Video Technology, Vol.9, No.6, pp 843-848.
[24] Su, S. L., Su, Y. C. & Huang, J. F. (2000) "Grey-based power control for DS-CDMA cellular mobile systems", IEEE Transactions on Vehicular Technology, Vol.49, No.6, pp 2081-2088.
[25] Jiang, B.C, Tasi, S. L & Wang, C. C. (2002) "Machine vision-based gray relational theory applied to IC marking inspection", IEEE Transactions on Semiconductor Manufacturing, Vol.15, No.4, pp 531-539.
[26] Luo, R. C, Chen, T. M & Su, K. L. (2001) "Target tracking using a hierarchical grey-fuzzy motion decision making method", IEEE Transactions on Systems, Man and Cybernetics, Part A, Vol.31, No.3, pp 179-186.
[27] Wang, Y. F. (2003) "On-demand forecasting of stock prices using a real-time predictor", IEEE Transactions on Knowledge and Data Engineering, Vol.15, No.4, pp 1033-1037.
[28] Huang, S. J, Huang, C. L. (2000) "Control of an inverted pendulum using grey prediction model", IEEE Transactions on Industry Applications, Vol.36, No.2, pp 452-458.
[29] Li, G, Ruhe, J, Emran, A. Al & Richter, M.M. (2007) "A flexible method for software effort estimation by analogy", Empirical Software Engineering, Vol.12, pp 65-106. [doi:10.1007/s10664-006-7552-4]

Authors

Geeta Nagpal

Geeta Nagpal, Ph D in Computer Science & Engineering from National Institute of Technology, Jalandhar, INDIA. She completed her Master's degree in Computer Science from Punjab Agricultural University, Ludhiana. She is presently working as Associate Professor in the Department of Computer Science and Engineering at National Institute of Technology, Jalandhar. Her research interests are Software Engineering, Databases and Data mining.

Prof. Moin Uddin

Prof. Moin Uddin, Pro Vice Chancellor, Delhi Technological University, Delhi ,INDIA. He obtained his B.Sc. Engineering and M.Sc. Engineering (Electrical) from AMU, Aligarh in 1972 and 1978 respectively. He obtained hid Ph. D degree from University of Roorkee, Roorkee in 1994. Before Joining as the Pro Vice Chancellor of Delhi Technological University, he was the Director of NIT, Jalandhar for five years. He has worked as Head Electrical Engineering Department and Dean Faculty of Engineering and Technology at Jamia Millia Islamia (Central University) New Delhi. He supervised 25 Ph. D thesis and more than 30 M.Tech dissertations. He has published more than 40 research papers in reputed journals and conferences. Prof. Moin Uddin holds membership of many professional bodies. He is a Senior Member of IEEE.

Dr. Arvinder Kaur

Dr. Arvinder Kaur, Professor, University School of IT, Guru Gobind Singh Indraprastha University, Delhi, India. She completed her masters degree in Computer Science from Thapar Institute of Engineering and Technology and Ph D from Guru Gobind Singh Indraprastha University, Delhi. Prior to joining the school, she worked with Dr. B. R. Ambedkar Regional Engineering College, Jalandhar and Thapar Institute of Engineering and Technology. Her research interests include Software Engineering, Object-Oriented Software Engineering, Software Metrics, Microprocessors, Operating Systems, Artificial Intelligence, and Computer networks. She is a lifetime member of ISTE and CSI. She is also a member of ACM. She has published 45 research papers in national and international journals and conferences. Her paper titled, "Analysis of object oriented Metrics"was published as a chapter in the book Innovations in Software Measurement (Shaker-Verlag, Aachen 2005).