A Hybrid Technique using Grey Relational Analysis and Regression for Software Effort Estimation using Feature Selection

Geeta Nagpal, Moin Uddin, Arvinder Kaur

Abstract- Software Estimation Techniques present an inclusive set of directives for software project developers, project managers and the management in order to produce more accurate estimates or predictions for future developments. The estimates also facilitate allocation of resources' for Software development. Estimations also smooth the process of re-planning, prioritizing, classification and reuse of the projects. Various estimation models are widely being used in the Industry as well for research purposes. Several comparative studies have been executed on them, but choosing the best technique is quite intricate. Estimation by Analogy(EbA) is the method of making estimations based on the outcome from k most analogous projects. The projects close in distance are potentially similar to the reference project from the repository of projects. This method has widely been accepted and is quite popular as it impersonates human beings inherent judgment skill by estimating with analogous projects. In this paper, Grey Relational Analysis(GRA) is used as the method for feature selection and also for locating the closest analogous projects to the reference project from the set of projects. The closest k projects are then used to build regression models. Regression techniques like Multiple Linear Regression, Stepwise Regression and Robust regression techniques are used to find the effort from the closest projects.

Index Terms—Estimation by Analogy, Feature Selection, Grey relational Analysis, Regression.

I. INTRODUCTION

Software Development is a creative process where in each person's productivity is different. Therefore, it is difficult to plan and estimate at the beginning as we know very less about the software being developed. As the process of development grows, we can analyze and get together deficient information and optimize the development process. Bohem and Valerdi[1] also stated that uncertainty level of effort estimation during early stage of software development is very high, while it decreases as the project progresses. This inherent uncertainty and imprecision of the software project impose challenges to software effort prediction. The software estimation process integrates approximating the size of the software product to be produced, estimating the effort required, developing preliminary project plan, and eventually, estimating on the entire cost of the project. Different software projects pose new challenges, as they are distinct from other projects. Software Developers and researchers are using different techniques and are more concerned about accurately predicting the effort of the software product being developed.

Even small enrichment in the prediction accuracy and validity are highly valued by researchers and software developers. Explorations are being made into alternative computationally intelligent hybrid software estimation model using analogy, case based reasoning, neural nets, fuzzy techniques, genetic algorithms, Bayesian networks etc. These explorations add to the ever retreating panorama of ideal resource estimation models, keeping the model development and evolution commune in a highly motivating and challenge driven state.

In this paper a new methodology based on integration of Grey Relational Analysis (GRA) and Regression techniques is proposed in order to overcome the challenges of uncertainty and imprecision. GRA is still at a nascent stage in the field of Software Estimation. It was first developed by Deng[2][3][4] in order to study the uncertainties that exist in the data. Now, it is being applied in varied fields such as decision making[5], transportation[6], manufacturing[6] and system control[5] etc. It is a problem solving process that is used to review the likeness between two tuples with same features. It is used to lessen the uncertainty in distance measure between two software projects. Though GRA and Regression have been used for software effort estimation, but hardly any research has been carried to exploit GRA and Regression techniques together as a hybrid model for estimation of software process. Shepperd,M., et al[7] expresses Estimation by analogy in an automated environment known as ANaloGy softwarE tooL (ANGEL) that supports the collection, storage and identification of the most analogous projects from the repository in order to estimate the cost and effort. It uses Euclidean distance as the distance measure to reduce the amount of computation involved. The research was carried out on six different datasets and it has outperformed the traditional algorithmic methods. Shepperd et al.[8] also validated nine different industrial datasets and concluded that in all cases analogy outperforms algorithmic models based on Step wise regression. On the other hand Angelis et al. [9] proposes the use of a statistical simulation procedure to improve upon Estimation by Analogy. Song et al.[10] predicts Software Effort with small data sets using GRA of Grey System Theory(GST) for feature subset selection and effort prediction. GST provides reliable analysis results. It utilizes the known data to set up an analysis model. This technique has been used to address problems in Image processing, mobile communication, machine vision, system control, stock price prediction.

Manuscript received on October, 2011

Geeta Nagpal,, Lecturer,CS Department, National Institute of Technology, Jalandhar, INDIA, 9888582299,sikkag@gmail.com

Moin Uddin, Pro Vice Chancellor, Delhi Technological University, Delhi ,INDIA, 9810553516, prof_moin@yahoo.com.

Arvinder Kaur, Associate Professor, University School of IT, GGSIPU, Delhi ,INDIA, 9810434395 , arvinderkaurtakkar@yahoo.com

The paper first focuses on feature subset selection as a process of identifying and removing redundant and irrelevant features for improving the prediction accuracy. They have proposed Grey Relational Analysis based on Software ProjeCt Effort Prediction (GRACE) including feature subset selection. Huang et al.[5] made software effort estimation based on Similarity distances. They have applied Genetic Algorithm to analogy based software effort estimation models. It is used to derive linear model from the similarity distances between pairs of projects for adjusting the reused effort. Li et al. [11] describes a new flexible method called AQUA which combines the key features from two known analogy based Estimation techniques: case based reasoning (CBR) and collaborative filtering(CF). The results have demonstrated better accuracy and broader applicability by combining techniques of CBR and CF with existing analogy-based effort estimation methods. Mittas et al.[12] uses a re sampling method in order to improve the estimation by analogy(EbA). They proposed the effect of iterated bagging on EbA and validated it using artificial and real data sets. Jorgenson and Shepperd [13] made a very systematic review, they considered 304 studies describing Research on Software Cost Estimations. According to them, roughly half of the studies used regression based estimation approach the most common parametric approach is the COCOMO model in which they have tried to focus on improvement or comparison with regression based estimation models. Hsu et al.[14] uses Weighted Grey Relational Analysis for Software Development and have proposed six weighted methods, namely, non weight, distance based weight, correlative weight, linear weight, non linear weight and maximal height to be integrated into GRA. Mitta's et al.[15] combined Regression and Estimation by Analogy in a Semi-parametric Model for Software Cost Estimation. Both methodologies by the acronym LSEbA were used on two data sets Abran and Robillard dataset and ISBSG dataset. The variables were partitioned into an LS and EbA subset. The results were improved by the utilization of this semi parametric model. The improvement was evaluated through various accuracy measures. Azzeh et al.[16] have used Estimation by Analogy based on the integration of Fuzzy set theory with Grey Relational Analysis(GRA).GRA has been used to reduce un certainty in the distance measure between two software projects for both continuous and categorical features. It has been used to determine the similarity between two data tuples.GRA is used to access the Grey Relational Grade(GRG) between the reference and comparative project. Pahariya et al.[17] in the paper "Computational Intelligence Hybrids applied to Software Cost Estimation have proposed Hybrid architectures involving Genetic Programming(GP) and Group Method for data handling(GMDH)". Mittas et al.[18] focused on parametric Regression Analysis and non parametric Estimation by analogy. In the paper, they have discussed the process of building a partially linear model and it achieves to incorporate both parametric and non parametric relationships into a simple parametric model. The remainder of this paper is organized as follows: In section 2, the Grey relational Analysis and Regression modeling techniques are discussed. Section 3, presents the proposed hybrid methodology, In section 4, we give the Experimental results of the methodology applied on four different public datasets. In section 5, the conclusions drawn from the results and some future directions for research.

II. MODELING TECHNIQUES

Estimation by Analogy

Estimation by Analogy, is a method that estimates a given software project with one or more projects that are similar to it from the historical set of projects. There should be a logical association between the given project and set of analogous projects. Various popular similarity measures used for measuring the distance from reference project to other projects are : Euclidean, Euclidean Square, City Block, Chebychev etc. Average or weighted average of the effort of the similar projects is used to compute the effort for the given project. Some of the other methods involved in finding the effort are Grey Relational Analysis, Machine Learning, Regression Techniques, Soft Computing methods and a combination of these, as shown in figure1.

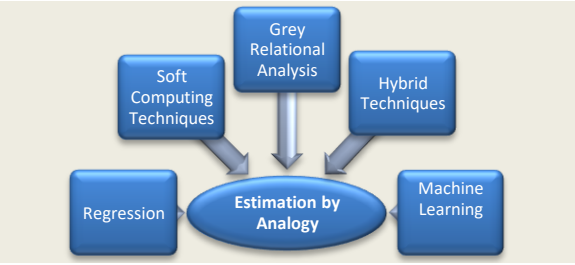

FIG 1: Estimation by Analogy Methods.

A. Grey Relational Analysis(GRA)

This is comparatively a novel technique in Software Estimations. It is a technique of Grey Systems Theory (GST) which was introduced by Professor Ju-Long Deng[2][3][4] and Wu [19][20][21]. The term ***"Grey"*** lies between "*Black"*(meaning no information) and "*White*"(meaning full information) and it indicates that the information is partially available. It is suitable to unascertained problems with poor information. GRA assists to work on such incomplete and unascertained information. GRA has the ability to learn from a small number of cases which is effective in the context of data-starvation. The magnetism of GRA is its flexibility to model complex nonlinear relationship. It utilizes the effort from historical projects of the same extent. Consider the objective series $\mathbf{x_i}$ **(i = 1, 2,...,n)** and the reference series x_0 ***(o =1,2,.....n).*** The three steps involved in the Wu's method are:

Grey Relational Grade by Wu's Method

Step 1: Data Processing: The first step is the standardization of the various attributes, so that every attribute has the same amount of influence, thus the data is made dimensionless, by using various techniques as Initial value processing ,average value processing, upper bound effectiveness, lower bound effectiveness or moderate effectiveness

a)Upper-bound effectiveness (i.e., larger-the-better)

$$x_i^*(k) = \frac{x_i(k) - min_i x_i(k)}{max_i x_i(k) - min_i x_i(k)}, \quad (1)$$

where $i=1,2,\ldots,m$ and $k=1,2,\ldots,n$.

b)Lower-bound effectiveness (i.e., smaller-the-better)

$$x_i^*(k) = \frac{max_i x_i(k) - x_i(k)}{max_i x_i(k) - min_i x_i(k)}, \quad (2)$$

where $i=1,2,\ldots,m$ and $k=1,2,\ldots,n$.

c)Moderate effectiveness (i.e., nominal-the-best)

$$x_i^*(k) = 1 - \frac{|x_i(k) - x_{ab}(k)|}{max\{max_i x_i(k) - x_{ab}(k),\ x_{ab}(k) - max_i x_i(k)\}} \quad (3)$$

where $i=1,2,\ldots,m$ and $k=1,2,\ldots,n$.

where $x_i(k)$ represents the value of the k_{th} attribute in the i_{th} series; $x_i^*(k)$ represents the modified grey relational generating of the k_{th} attribute in the i_{th} series; $max_i x_i(k)$ represents the maximum of the k_{th} attribute in all series; $min_i x_i(k)$ represents the minimum of the k_{th} attribute in all series, and $x_{ab}(k)$ is the objective value of the k_{th} attribute.

Step 2: Difference Series: Difference series is calculated as:

$$\Delta_{ij} = \left| x_i^*(k) - x_j^*(k) \right| \quad (4)$$

where $i, j=1,2,\ldots,n$.

Step 3: Globalized Modified Grey Relational Grade: The Globalized Modified Grey Relational Grade of the i_{th} series and the j_{th} series is given by:

$$\Gamma_{ij} = \frac{\Delta_{min} + \Delta_{max}}{\bar{\Delta}_{ij} + \Delta_{max}}, \quad i, j = 1,2,\ldots,m. \quad (5)$$

where, $\Delta_{min} = min_i\, min_j\, min_k |x_i(k) - x_j(k)|$,

$\Delta_{max} = max_i\, max_j\, max_k |x_i(k) - x_j(k)|$,

and $\bar{\Delta} = \frac{1}{n}\sum_{i=1}^{n} \Delta_{ij}(k)$.

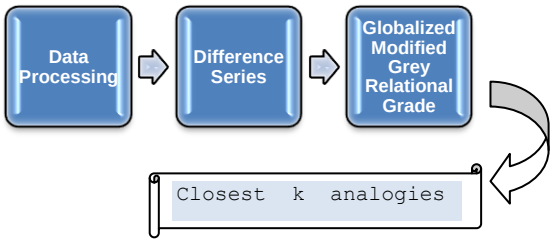

FIG 2: GRA PROCESS

B. Regression Techniques

Regression analysis is a statistical techniques for modelling and analysis of variables. Regression is used to study the relationship that exists between dependent variable and one or more independent variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable whose value influences or is used for prediction is called Independent variable. It is also called regressor , predictor or explanatory variable.

$$Y = a + bx + e \quad (6)$$

Y is the dependent variable
a is the intercept
b is the regression coefficient
x is the predictor variable

Multiple Linear Regression

A multiple linear regression model generalizes the simple linear regression model by allowing the response variable to depend on more than one explanatory variable As in the case of simple linear regression models, the principle of least squares is used to fit the regression line. According to the principle of least squares the `best fitting' line is the line which minimizes the deviations of the observed data away from the line. This line is called the least squares line. The regression parameters for the least squares line, the least squares estimates, are estimates of the unknown regression parameters in the model. The coefficient of multiple determination is a measure of how well the fitted line describes the variation in the data. The equation for Multiple linear regression is given by:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots\ldots\ldots \beta_k\, x_{i,k} + \varepsilon_i \quad (7)$$

where, Y_i is the response variable or dependent variable where as, $x_1, x_2, x_3 \ldots\ldots\ldots x_k$ are k independent or explanatory variables. β_o is the y intercept, β_1 , β_2 are the slope of y with independent variables $x_{i,1}$ and $x_{i,2}$.

Robust Regression

Robust Regression is a type of regression technique which prevails over limitations of ordinary least square. Ordinary least square estimates are extremely non-robust to outliers. Outliers are those observations in the dataset that do not follow the prototype of the other observations. These outliers can inefficiently influence the whole process of fitting. They should be either treated or detached from analysis. The purpose of robust regression is to provide resistant results even in the presence of outliers. They are useful when the error is not normally distributed. Robust Regression down weights the effects of the observations with large residual error. Various Robust Regression Techniques exist like M Estimators, Least Trimmed Squares(LTS), S Estimation etc. In this paper various M estimators like 'huber-estimator', 'bisquare-estimator','Cauchy-estimator', logistic-estimator', 'andrew-estimator', 'fair-estimator', 'talwar-estimator' and 'welsch-estimator' etc have been used with their different weight functions for predicting the software effort of the projects. M Estimation was introduced by Huber[22][23] as the most general and widespread method of robust regression. M-estimators are also called Maximum-likelihood estimation as it minimizes a weighted sum of residuals.

The steps involved are:

Step 1:In the first iteration, each observation is allocated equal weight and model coefficients are estimated using ordinary least squares(OLS).

Step 2:In the second step, OLS residuals are used to find weights. The observation with larger residual is assigned lower weight, so that observations farther from model predictions in the earlier iteration are given lesser weight.

Step 3:In the third iteration, the new model parameters and the residuals are recomputed using weighted least squares(WLS).

Step 4:In step 4, new weights as per step 2 are found and the procedure continues until the values of the parameter estimates converge within a specified tolerance.

For fitting of robust regression on the k closest projects MATLAB [24] is used. It uses reweighted least square with different weighting functions.

b=robustfit (X, y, wfun, tune) (8)

The function returns a *p*-by-1 vector b of parameter estimates for a robust multi linear regression of the responses in y on the predictors in X[24]. X is an *n*-by-*p* matrix of *p* predictors at each of *n* observations. y is an *n*-by-1 vector of observed responses. It specifies the weighting function *wfun* , and a tuning constant *tune*. Tune is chosen to provide higher efficiency,

smaller the value of tuning constant the more resistant it is to the outliers as it increases the down weight assigned to large residuals; on the contrary, increasing the tune constant decreases the down weight. The following weighting function have been used with all the projects in all the four datasets[24].

TABLE I. Robust Regression Functions

Weight Function	Equation	Default Tuning Constant
'andrew'	$w = (abs(r) < p_i).* \sin(r)./r$	1.339
'bisquare'	$w = (abs(r) < 1).*(1 - r.\hat{}2).\hat{}2$	4.685
'cauchy'	$w = 1./(1 + r.\hat{}2)$	2.385
'fair'	$w = 1./(1 + abs(r))$	1.400
'huber'	$w = 1./\max(1, abs(r))$	1.345
'logistic'	$w = \tan(r)./r$	1.205
'talwar'	$w = 1 * (abs(r) < 1)$	2.795
'welsch'	$w = \exp(-(r.^2))$	2.985

The value of r in the weight function is

$$r = \frac{resid}{(tune * s * sqrt(1-h))} \quad (9)$$

where, *resid* is the vector of residuals from the previous iteration, *h* is the vector of leverage values from a least square fit, and *s* is the estimate of the standard deviation of the error term.

Stepwise Regression

It is methodical method for adding and removing terms based on their statistical importance .The main approaches are:

- Forward selection: This method which involves starting with no variables in the model, trying out the variables one by one and including them if they are 'statistically important.
- Backward elimination, which involves starting with all candidate variables and testing them one by one for statistical significance, deleting any that are not important.
- Combination of the above methods, testing at each stage for variables to be included or excluded.

C. Modelling Process:

The process of modelling is a process which involves a number of steps as shown in the Figure 3:

Step 1: Data Acquisition: The projects are collected from various repositories like PROMISE, NASA, IFPUG etc. In this paper datasets have been taken from the PROMISE repository.

Step 2: Data Pre-processing: The data is pre-processed and cleaned before being put to any kind of analysis.

Step 3: Feature subset selection: There are number of features concerned in prediction of effort , but some may have higher degree of influence on the output than the others. Those which have higher influence are considered more important . Thus a optimal feature subset for effort prediction has to be generated. In this paper Grey Relational Analysis(GRA) has been used for Feature Selection.

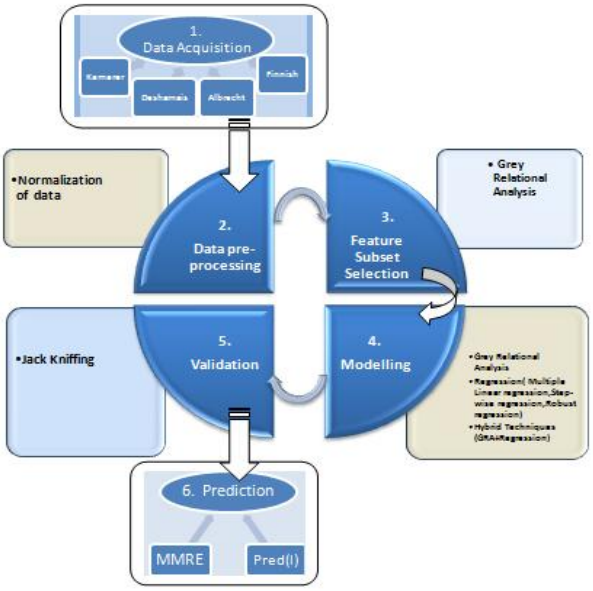

FIG 3. Process of Modelling for Software Effort Estimation

Step 4: Modeling Techniques: Various modelling techniques are available as discussed earlier. In this work, Grey relational analysis has been used for finding the effort from k nearest projects. Regression techniques like Multiple Linear Regression ,Robust Regression techniques and Stepwise Regression are performed iteratively for varying number of closest projects. The projects that have the highest value on Grey Relational Grade gets the greatest opportunity to contribute in the final estimate.

Step 5: Validation : Various methods of validation like hold out method, random sub sampling, 10 fold validation and leave one out cross validation etc exist. In this work Leave one out cross validation for empirical evaluation has been used. This method is also called Jack-knifing.

Step 6:Prediction : Various criteria exists for the evaluation and prediction of software effort like Magnitude of relative Error(MRE), Mean Magnitude of Relative Error(MMRE),Mean squared error (MSE),Pred(l) etc. In this paper, MMRE and Pred(l) have been used.

TABLE II Various Performance Measures

Performance Measures	Formula
RE (Relative Error)	$actual_i - estimated_i \,/\, actual_i$
MRE (Magnitude of Relative Error)	$(\lvert actual_i - estimated_i \rvert) \,/\, actual_i$
MMRE (Mean MRE)	$\frac{1}{n}\sum_{i=1}^{n} MRE_i$
MSE (Mean Squared Error)	$\frac{1}{n} * \sum_{i=1}^{n} (actual_i - estimated_i)^2$
MdMRE(Median of MRE's)	$median_i \ (MRE_i)$
MSRE (Mean Squared Relative Error)	$\frac{1}{n} * \sum_{i=1}^{n} (MRE_i)^2$
SD (Standard Deviation of RE)	$\sqrt{n * \sum_{i=1}^{n} RE_i^2 - (\sum_{i=1}^{n} RE_i)^2 / n * (n-1)}$
Pred(l)	k/n*100

III. A PROPOSED METHODOLOGY

A. Modeling using GRA and fusion of Grey Relational Effort Analysis Technique with Regression Methods (GREAT_RM):

Step 1: Finding k closest projects by Grey Relational Analysis GRA:

As per the steps given in Sec II A, the Globalized Grey Relational Grade for each project is generated which lies between 0 and 1. The projects that have the highest value on Grey Relational Grade gets the greatest opportunity to contribute in the final estimate

Step 2: Effort Prediction : Effort for GRA is the simple aggregation of k most influential projects.

$$\hat{\varepsilon} = \sum_{i=1}^{k} w_i * \varepsilon_i \quad (10)$$

where weight w_i is given by,

$$w_i = \frac{\Gamma(x_0, x_1)}{\sum_{j=1}^{k} \Gamma(x_0, x_j)} \quad (11)$$

Step 3: Effort Prediction by Regression applied on GRA:

In this step, effort estimate for a given project is calculated based on the GREAT_RM technique. Various Regression techniques like Multiple Linear Regression, Robust Regression and Step wise Regression Techniques are applied to the k closest projects based on Globalized Grey Relational Grade. The value of k is not predetermined, the value of k will vary with each iteration and the final value of k would be the one with lowest error.

Step 4: Model Generation:

A model is generated using the lowest Magnitude of Relative Error(MRE) as the criteria for all the projects. The Figure 4 clearly shows the flow of data. The model generated can be used for effort generation of any similar project.

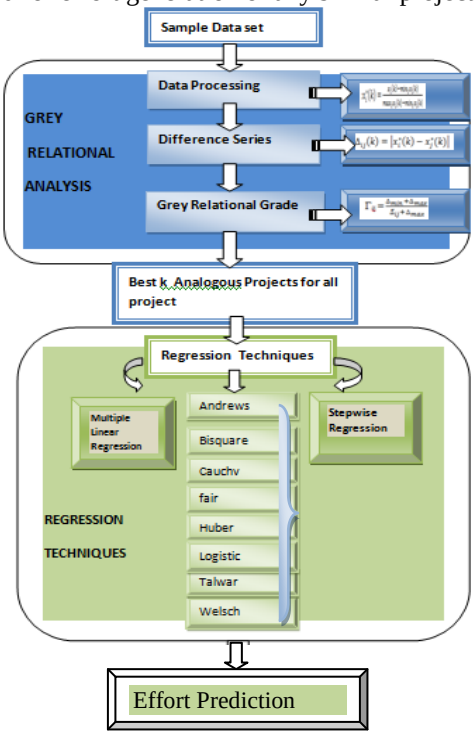

FIG 4. Flow Diagram of *GREAT_RM* Technique.

.

B. Modeling using GREAT_RM+ (GREAT_RM with Feature Selection):

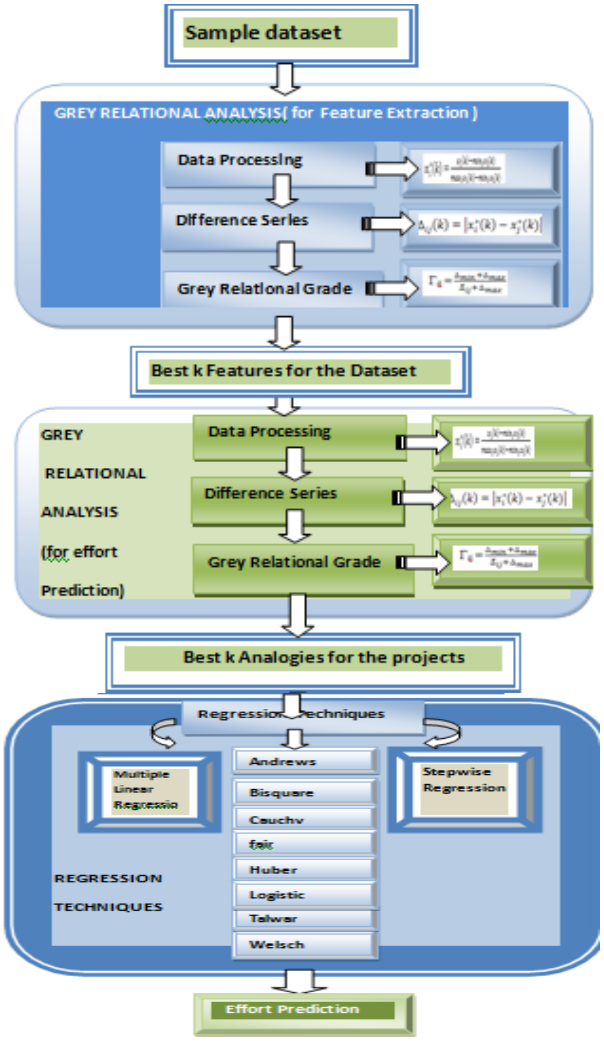

FIG 5. Process of *GREAT_RM* + Technique with feature Selection

Step 1: Feature Extraction by Grey Relational Analysis[10]

1: Construction of data: The columns in the dataset are treated as series . The Effort series $x_e = \{e_1, e_2, e_3 \ldots\ldots e_n\}$ is taken as the reference series and the attribute columns are regarded as objective series.

2: Normalization: Each data series is normalized inorder to have the same degree of influence on the dependent variable.Thus they are made dimensionless.

3: Generation of Grey Relational Grade:Grey Relational grade is calculated for each series.

4: Feature Selection : Features with higher value of GRG encompass the most favourable feature subset.Best results can be generated by varying the number of features.

Step 2: Effort Estimation by Grey Relational Analysis as given in section III A

Step 3: Regression Techniques applied on best k analogous projects as given in section II B

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Step 4: Model building using lowest MMRE for projects. Figure 5, clearly shows the flow of data. The model generated can be used for effort generation of any similar project.

IV. EXPERIMENTAL METHODS AND RESULTS

A. Data Sources

The data used in the present study comes from PROMISE repository . This hybrid technique ***GREAT_RM***$^+$ was tested on four public datasets namely Finnish, Albrecht , Kemerer and Desharnais.

Finnish Data set: This data set consists of 38 projects and 9 attributes namely "Project Id", "hw", "at" "FP", "Co", "prod", "Insize" ,"Ineff" and "dev.eff.hrs".

Albrecht Data set: It contains 24 projects and 9 attributes which include "Input Count", "Output Count", "Inquiry Count", "File count", "FPAdj", "RAWFP count ", "AdjFP" and "Effort".

Kemerer Data set: The Kemerer dataset has 15 Software Projects and 8 attributes namely "Project ID", "Language", "Hardware", "Duration", "KSLOC", "AdjFP", "RAWFP" and "Effort".

Desharnais Data set: The desharnais dataset comprises of 81 projects from Canadian Software houses. It has 11 attributes namely "Project Id", "Team Experience", "Manager Experience", "Year End", "Length", "Transactions", "Entities", "Points Non adjustable", "Adjustment Factor", "Points Adjust", "Language" and "Effort.

B. Validation Method

Wu's Grey Relational Grade calculation method has been used for anticipating the closest projects. The datasets available are small, therefore instead of using hold out method or 10 fold validation, jack knife method is used for validation. In the Process of Estimation by the GREAT_RM$^+$ Technology, various regression techniques like Multiple Linear Regression, Stepwise Regression and Robust Regression techniques have been applied. The value of k is not predetermined, it varies with each iteration. The final value of k is the one with lowest MMRE for all projects. For evaluation of performance of the proposed GREAT-RM$^+$,we have compared the results with the results obtained from GREAT_RM(without Feature Extraction) and also the results obtained by GRACE[10].

C. Evaluation Criteria

In this paper, Mean Magnitude of Relative error (MMRE) and Pred(30) are used as the criteria for evaluating software effort prediction. Relative Error is the absolute error in the observation divided by its true value. The MRE is a percentage of the actual effort for the project, whereas Mean MRE is the aggregation of all MRE to the number of projects. Pred (l) is the complementary criteria of MRE. l is the number of projects, where MRE_i <=l % A low score on MRE and MMRE whereas a high value on pred (l) entail better accuracy.

D. Experimental Results

FINNISH DATASET: The results of the GREAT-RM$^+$ techniques applied on Finnish dataset are shown in the table III. The best MMRE achieved was 14.38 with GRA and Robust Regression technique. The result is better than result obtained using GREAT-RM (without Feature extraction) with MMRE of 34.76. The results are much better than using only GRA where the MMRE was 62.3833. The results obtained are very encouraging. Figure 6 and 7 clearly demonstrate the improvement in result on using the GREAT-RM$^+$ methodology.

TABLE III: Results of Finnish dataset (with and without Feature Extraction)

Data set	MMRE(%)	MMRE(%) of Grey Relational Effort Analysis Techniques with Regression Methods (GREAT_RM)									
	GRA	GRA and MLR	GRA and Andrew	GRA and Bisquare	GRA and Cauchy	GRA and Fair	GRA and Huber	GRA and Logistic	GRA and Talwar	GRA and Welsch'	GRA and Stepwi
KEMERER DATASET	63.691	51.774	48.75	48.749	48.782	48.76	48.679	48.686	48.793	48.686	57.8
KEMERER DATASET(with Feature Extraction)	62.047	52.834	48.428	48.426	48.467	48.439	48.831	48.344	48.494	48.344	52.678

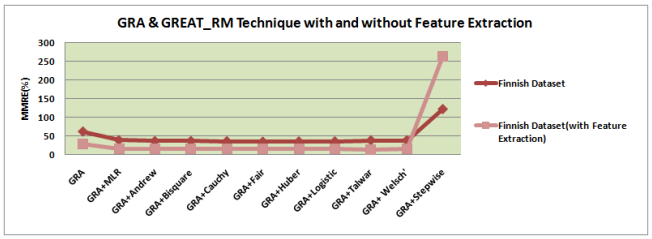

FIG 6. Line Graph for Finnish Dataset

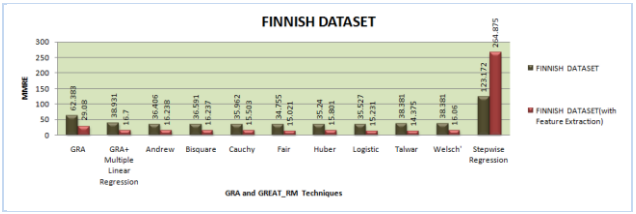

FIG 7. Bar graph showing the comparison of Finnish Dataset

ALBRECHT DATA SET: Albrecht dataset consists of 24 projects. After the GREAT_RM$^+$ technique is applied to it ,some vital results were obtained. The best MMRE for 22 analogies was obtained as 33.15 which is considerably better than MMRE of 59.3804 achieved by GREAT_RM without Feature Extraction and also better than using only GRA having MMRE of 78.57. The results obtained as shown in Table IV, Figure 8 and 9 which are quite impressive, and certainly a great improvement over GRACE [10]with MMRE of 60.25.

TABLE IV. Results of Albrecht Dataset

Data set	MMRE(%	MMRE(%) of Grey Relational Effort Analysis Techniques with Regression Methods (GREAT_RM)									
	GRA	GRA and MLR	GRA and Andre	GRA and Bisqu	GRA and Cau	GRA and Fair	GRA and Hub	GRA and Log	GRA and Talw	GRA and We	GRA and Ste
ALBRECHT DATASET	78.57	68.441	63.679	63.808	62.803	62.817	59.38	62.516	64.902	61.711	69.49
ALBRECHT DATASET(with Feature Extraction)	37.3	35.149	33.508	33.506	33.459	33.678	33.469	33.436	33.867	33.497	47.94

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

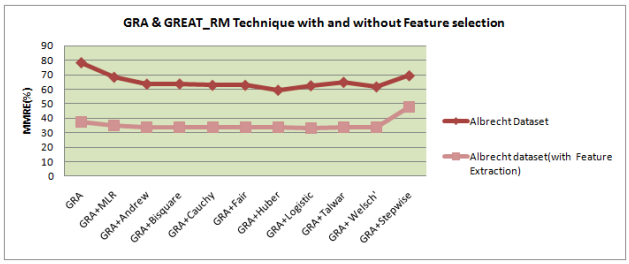

FIG 8. Comparison of Results using a Line graph

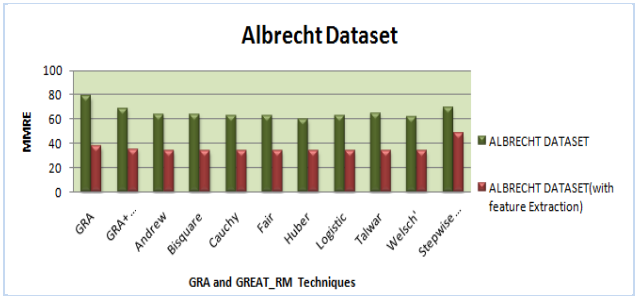

FIG 9. Bar graph showing the comparison of Albrecht Dataset

KEMERER DATASET: The data set has achieved MMRE of 48.344 using the GREAT_RM^{+} technique which is an improvement over GREAT_RM without Feature Extraction with MMRE of 48.6792. The method is certainly better than using only GRA with MMRE of 63.691 and GRACE[10] with MMRE of 58.83. Table V and Figure 10 show the results obtained for Kemerer Data set.

TABLE V. Results of Kemerer Dataset

Data set	MMRE(%)	MMRE(%) of Grey Relational Effort Analysis Techniques with Regression Methods (GREAT_RM)										
	GRA	GRA and MLR	GRA and Andrew	GRA and Bisquare	GRA and Cauchy	GRA and Fair	GRA and Huber	GRA and Logistic	GRA and Talwar	GRA and Welsch'	GRA and Stepwise	
KEMERER DATASET	63.691	51.774	48.75	48.749	48.782	48.76	48.679	48.686	48.793	48.686	57.8	
KEMERER DATASET(with Feature Extraction)	62.047	52.834	48.428	48.426	48.467	48.439	48.831	48.344	48.494	48.344	52.678	

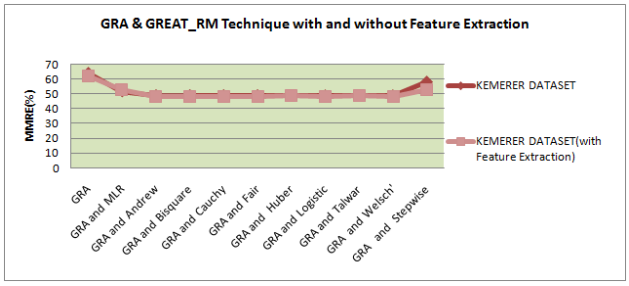

FIG 10. Graph showing the comparison of various techniques applied on Kemerer Dataset (with and without Feature Extraction)

DESHARNAIS DATASET: Desharanis Dataset consisting of 81 projects. The MMRE of 42.24 was achieved on using GREAT_RM^{+} technique which is better than MMRE of 43.556 obtained using GREAT_RM. The results are significant improvement over GRA with MMRE of 53.015 and also GRACE[10] with MMRE of 49.83. Results obtained are shown in Table VI and Figure 11.

TABLE VI Results of Desharnais Dataset

Data set	MMRE(%)	MMRE(%) of Grey Relational Effort Analysis Techniques with Regression Methods (GREAT_RM)									
	GRA	GRA and MLR	GRA and Andrew	GRA and Bisquare	GRA and Cauchy	GRA and Fair	GRA and Huber	GRA and Logistic	GRA and Talwar	GRA and Welsch'	GRA and Stepwise
DESHARNAIS DATASET	53.015	43.069	43.663	43.609	43.609	44.154	44.79	43.556	43.907	43.712	3293.091
DESHARNAIS DATASET(with Feature Extraction)	57.111	45.818	43.042	43.058	42.238	42.516	42.504	42.303	42.164	43.399	420.292

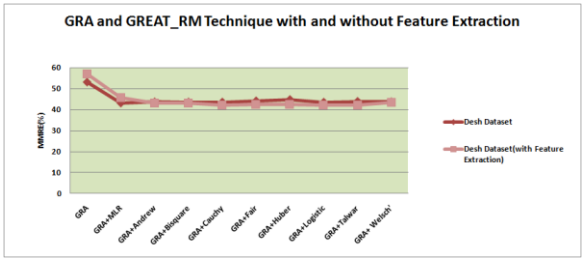

FIG 11. Comparison using Line graph for Desharnais Dataset

The line graph of the two techniques are shown below in Figure 12 and Figure 13. The results have been compared with GRA and GRACE[10](Figure 14)

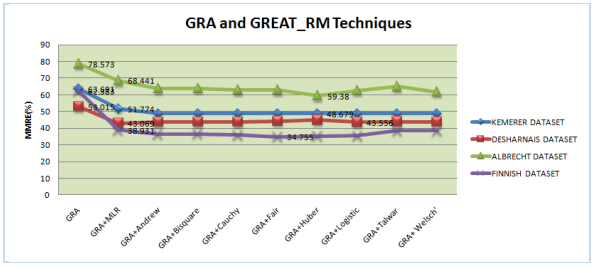

FIG 12. Result of GREAT_RM Technique without Feature Selection

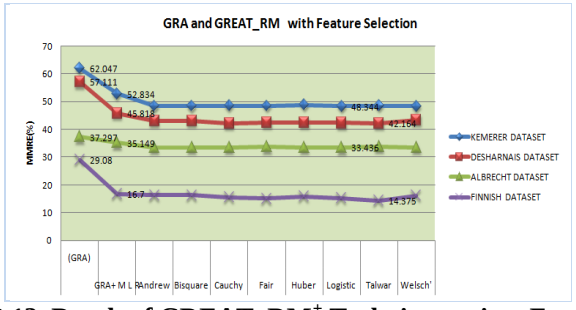

FIG 13. Result of GREAT_RM^{+} Technique using Feature Selection

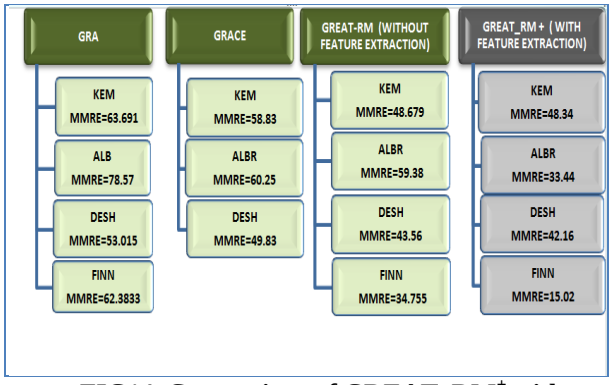

FIG14. Comparison of GREAT_RM^{+} with GREAT_RM(without Feature Extraction) and GRACE and GRA

V. CONCLUSION AND FUTURE SCOPE

"Modeling a hybrid technique using Grey Relational Effort Analysis Technique with Regression Methods including Feature Selection (GREAT_RM$^+$) " has achieved admirable results. The results obtained with Feature Extraction have been commendable and are compared to results obtained without Feature Extraction and with only GRA and GRACE[10]. The proposed method has shown striking results in Finnish and Albrecht datasets. The empirical evaluation have revealed that the GREAT_RM$^+$ techniques can remarkably improve the estimation process. The proposed method has outperformed some well known estimation techniques. The model GREAT_RM$^+$ can be used for early stage estimation where the data is uncertain. This is can further be applied on some other large datasets with different validation methods.

REFERENCES

1. B.W. Boehm, and R.Valerdi. "Achievements and Challenges in Software Resources Estimation", CSSE Tech Report,2005
2. J L Deng "Control problems of grey system". System and Control Letters, 1:288–94,1982
3. J Deng "Introduction to grey system theory". Journal of Grey System 1:1–24,1989
4. J Deng "Grey information space". Journal of Grey System 1:103–117,1989
5. C J Hsu, C Y Huang "Improving Effort Estimation Accuracy by Weighted Grey relational Analysis During Software development". 14th Asia-Pacific Software Engineering Conference,534-541,2007
6. S J Huang , N H Chiu , L W Chen "Integration of the grey relational analysis with genetic algorithm for software effort estimation". European Journal of operational and research 188:898-909,2007
7. M. Shepperd, C. Schofield,B. Kitchenham. "Effort Estimation using Analogy". In Proceedings of the 18th International Conference on Software Engineering,170-178,1996
8. M. Shepperd,C. Schofield."Estimating Software Project Effort Using Analogies". IEEE Transactions on Software Engineering,23(12):736-743,1997
9. L. Angelis,I. Stamelos."A simulation tool for efficient analogy based cost estimation". Empirical Software Engineering, 5 :35–68,2000
10. Q. Song, M. Shepperd, C. Mair,"Using Grey Relational Analysis to Predict Software Effort with Small Data Sets". Proceedings of the 11th International Symposium on Software Metrics (METRICS'05), 35-45,2005
11. J. Li, G. Ruhe, A. Al-Emran, M. M. Richter. "A flexible method for software effort estimation by analogy", Empirical Software Engineering,12:65–106,2007
12. N.Mittas, M.Athanasiades, Angelis. "Improving analogy-based software cost estimation by a re sampling Method". Journal of Information & software technology,50,2008
13. J. Jorgenson , Shepperd. "A Systematic Review of Software Development Cost Estimation Studies". IEEE Transactions on Software Engineering,33(1),2007
14. C. J. Hsu and C. Y. Huang, "Improving Effort Estimation Accuracy by Weighted Grey relational Analysis During Software development". 14th Asia-Pacific Software Engineering Conference,534-541,2007
15. N.Mittas, L.Angelis,"LSEbA: least squares regression and estimation by analogy in a semi-parametric model for software cost estimation". Empirical Software Engineering,15(5):523-555,2010
16. M. Azzeh, et al., "Analogy-based software effort estimation using Fuzzy numbers". Journal of Systems and Software,84:270-284,2011
17. J.S.Pahariya, V. Ravi, M. Carr, M.Vasu, "Computational Intelligence Hybrids Applied to Software Cost Estimation". International Journal of Computer Information Systems and Industrial Management Applications, 2:104-112,2010
18. N.Mittas, L.Angelis."Comparing cost prediction models by resampling techniques", Journal of Systems and Software, 81:616–632,2008
19. J. H. Wu, C. B. Chen. "An alternative form for Grey Relational Grade". Journal of Grey System,11(1):7–11,1999
20. K.H.Hsia, J.H.Wu. "A study on the data preprocessing in Grey Relation Analysis". Journal of Chinese Grey System,1:47-53,1998
21. J.H.Wu, M.L.You and K.L.Wen. "A Modified Grey Relational Analysis",The Journal of Grey System,11(3):287-292,1999
22. P.J. Huber, Robust regression: Asymptotics, conjectures and Monte Carlo, The Annals of Statistics, 1, 799–821, 1981.
23. P.J. Huber, "Robust Estimation of a Location Parameter". Annals of Mathematical Statistics, 35:73–101, 1964
24. MATLAB®Documentation, http://www.mathworks.com/help/techdoc/

AUTHORS PROFILE

Geeta Nagpal, pursuing Ph D programme in the Department of Computer Science and Engineering, NIT, Jalandhar, INDIA She did her Master's degree in Computer Science from Punjab Agricultural University, Ludhiana. She is presently working as Lecturer 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, 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. He has worked as Head Electrical Engineering Department and Dean Faculty of Engineering and Technology at Jamia Millia Islamia (Central University) NewDelhi. He supervised 14 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, Associate Professor, University School of IT, Guru Gobind Singh Indraprastha University,Delhi.India and her master's degree in computer science from Thapar Institute of Engineering and Technology. 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 also a lifetime member of ISTE and CSI. She is also a member of ISTE,CSI,and ACM. Kaur has published 45 research papers in national and international journals and conferences. Her paper titled ''Analysis ofobject oriented Metrics'' was published as a chapter in the book Innovations in Software Measurement (Shaker-Verlag, Aachen 2005).