Analogy Based Software Project Effort Estimation Using Projects Clustering

M.Kowsalya[*], H.OormilaDevi[*], N.ShivaKumar[**]

[*] *Computer Science and Engineering Department, Thiagarajar College Of Engineering , Madurai*
[**] *Assistant Professor , Computer Science and Engineering Department, Thiagarajar College Of Engineering , Madurai*

Abstract— Software effort estimation is the methodology predicting the amount of effort required to develop or maintain the particular software. The Software effort estimation is task which is done in the requirement engineering phase where a requirement is taken and the effort needed for that particular requirement is found. The accuracy value of the effort must be very high because it may lead to the financial loss or the loss of reputation of the company. K-Means clustering algorithm is one of the machine learning oriented techniques. The clustering of the dataset is done using the K-Means clustering algorithm to estimate the effort accurately and efficiently. After clustering, the analogy based estimation technique is used. The non-algorithmic model like the analogy based effort estimation performs better than the algorithmic model like COCOMO. The effort estimated using the analogy based effort estimation coincides mostly with the actual effort obtained from the software effort estimation dataset. The results of the effort estimation were analyzed using the Magnitude of Relative Error (MRE) and Mean Magnitude Relative Error (MMRE) to prove its accuracy.

Index Terms - Analogy, Clustering, Effort, Estimation.

I. INTRODUCTION

Software effort estimation is one of the major and the important task in the software engineering. It is used for calculating the amount of effort required for the development and the management of the software project. It is one of the challenging task because whenever we estimate the effort for a particular software project we overestimate or underestimate the actual effort required. Resources are very limited in amount so a careful estimation of the effort must be done. Estimates of effort is usually measured in terms of person-months for a software project. Effort estimates must be very accurate because it may lead to financial loss or loss of reputation of the organization. Various algorithmic and non-algorithmic methods are used to estimate the effort of a particular software project.

The ways of categorizing estimation approaches, are Expert estimation where the effort is calculated using the judgmental processes of the experts in the organization. The experience of the experts is used for the effort estimation. Second is the Formal estimation model or the algorithmic model where the effort estimation is based on mechanical processes i.e., the mathematical and predefined formula is used. Some of the formal estimation model are COCOMO, Use case method etc., Third is the Learning Oriented estimation or the non-algorithmic model where the effort estimation is based on a machine learning techniques.

One of the machine learning oriented technique which is useful for the efficient effort estimation is the K-Means Clustering Algorithm. K-Means algorithm is one of the non-hierarchical algorithm for clustering. K-Means clustering algorithm aims to partition n dataset observations into k number of clusters. Since the accuracy value of the effort is very high we have chosen this algorithm for the effort estimation and the effort value found by this algorithm mostly coincides with the actual effort value. Learning oriented technique has the ability to learn from the previous historical data given and from that learned technique, the effort value can be easily measured accurately. After the clustering of the dataset, analogy based estimation technique is followed to estimate the effort of the software project accurately. In the analogy based software project effort estimation, the comparison of the current project with the historical project is done and as a result similar projects to the current project is obtained, using which the effort value is estimated.

The performance indicators used are the Magnitude of Relative Error (MRE) and Mean Magnitude of Relative Error (MMRE).These values must be as minimum as possible to make sure that the effort estimated is accurate and it mostly coincides with the actual effort given in the dataset.

Remaining sections of the paper are organized as follows:
Section II describes the background of the work i.e. Analogy based software project effort estimation, K-Means Clustering and evaluation criteria. Section III describes the related work in the field of software estimation. Section IV shows the combined model of K-Means clustering and analogy based estimation technique which is proposed in this paper. Section V describes the case study that is performed for result analysis. At the end of this paper, section VI shows the conclusion of the paper.

II. BACKGROUND

This section describes the various terms which are the core part of this paper such as k-means clustering, analogy based estimation technique and Evaluation criteria.

A. K-Means Clustering

K-means was introduced by James MacQueen in 1967. K-Means algorithm is one of the non-hierarchical algorithm for clustering .K-Means clustering algorithm aims to partition n dataset observations into k number of clusters.

B. Evaluation criteria

Mean Relative Error (MRE) :

MRE computes the percentage of error between actual and estimated effort for each reference project.

$$MRE = (Actual_i - Estimated_i)/ Actual_i$$

Mean Magnitude Relative Error (MMRE) :

MMRE calculates the average of MRE over all referenced projects.

$$MMRE = \frac{1}{n}\sum MREi$$

C. Analogy Based Effort Estimation

Analogy based estimation technique is followed to estimate the effort of the software project accurately. In the analogy based software project effort estimation, the comparison of the current project with the historical project is done and as a result similar projects to the current project is obtained, using which the effort value is estimated. Similarities are found by calculating the Euclidean distance in an n-dimensional space where each dimension represents a variable.

III. Related Work

Clustering is one of the vital data processing task .It has been used extensively by variety of researchers for various application areas like finding similarities in pictures, text information etc., K-Means algorithm is one in all the popular clustering algorithm [2]. A new hybrid toolbox based on soft computing techniques for effort estimation is introduced. Particle swarm optimization and cluster analysis has been enforced to perform economical estimation of effort values with intelligence [8]. The initial cluster centre is chosen by using particle swarms clustering algorithm underneath default variety of cluster, then optimizes the cluster, and last carries out cluster merging supported multiclass merging condition, so as to get the simplest cluster results [7]. This analysis uses some computing intelligence techniques, like Pearson product-moment correlation coefficient method and unidirectional ANOVA methodology to pick key factors, and K-Means cluster algorithmic rule to cluster the dataset and then estimate the software project effort [9]. A new methodology on Quad-tree and K-means algorithms which supports data processing ideas is outlined to assist the error between the actual effort and the estimated effort, within the initial stages of the project [5].

Analogy-based effort estimation (ABE) one in all the best and accurate ways for software project effort estimation .It has the outstanding performance and capability of handling noisy datasets. The effort worth is found accurately by doing clustering of the dataset using the K-Means cluster then analogy based effort estimation is done. The dataset used here is the ISBSG software effort estimation dataset. It has 17 attributes and 93 instances. The Performance indicators used here is the MRE and the MMRE values which gives the accuracy of the actual and the estimated effort values [1].The development effort is calculable by a comparison method during which the similar projects like a new project is chosen. Using the chosen projects, effort is then estimated for the new project. Due to simplicity and estimation capability, ABE has been extensively employed in terms of software development effort estimation.

IV. PROPOSED WORK

In the proposed work, initially the K-Means Clustering Algorithm is done to cluster the dataset into k number of clusters. The number of clusters into which the dataset is to be divided is found using the elbow plot. Once the clustering is completed the effort of the software project is found using the Analogy based effort estimation model.

A. Finding number of clusters

Finding the number of clusters in the K-Means clustering is always an issue. The number of clusters, k into which the dataset is to be divided is found using the elbow plot method. In the elbow plot method, the sum of squared error (SSE) value is used. The value of k is found such that it gives a better modeling of the data. The optimal number of clusters can be identified using the following steps :

Input: Dataset

1. Initially, compute the K-Means Clustering algorithm for the varying values of k i.e., from 2 to 10.
2. For each values of k, find the SSE value.
3. Plot the curve for SSE and the number of cluster k values.
4. Find the knee point, which gives the number of clusters, k.

Output: Number of clusters

B. K-Means Clustering

K-Means clustering algorithm aims to partition n dataset observations into k number of clusters.The K-means clustering algorithm is as follows

Input: Number of clusters, Dataset

1. Choose k objects from data sets as the centroids
2. Select the data points and find the Euclidean distance
3. Assign the data points to the cluster which has the minimum value
4. ReCalculate the centroid
5. If any changes in centroid redo the steps.

Output: k clusters

C. Analogy Based Effort Estimation

After the clustering of the dataset, analogy based estimation technique is followed to estimate the effort of the software

project accurately. In the analogy based software project effort estimation, the comparison of the current project with the historical project is done and as a result similar projects to the current project is obtained, using which the effort value is estimated. The project for which the effort is to estimated is assigned to a particular cluster and from that particular cluster the solution function is applied to find the effort. Similarities are found by calculating the Euclidean distance in a n-dimensional space where each dimension represents a variable. The steps involved in the analogy based effort estimation technique are :

Input: Clustered Dataset

1. Collect the Historical data set.
2. Using the Similarity function in the clustered dataset find the historical projects similar to the current project.
3. Find the final effort value using the Solution function.

Output: Software Effort

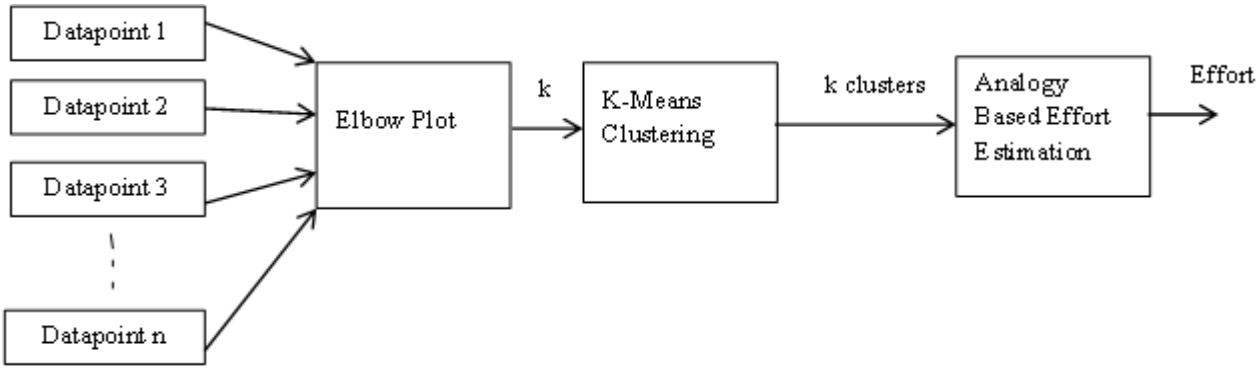

Figure 1: *Combined K-Means and Analogy Based Effort Estimation Model*

V. Results

In this proposed method, PROMISE NASA software effort estimation project data is considered which has 17 attributes. The dataset is divided as 80% for the training dataset and the remaining 20% as the testing dataset. Initially the dataset is divided into k number of clusters using the K-Means Clustering algorithm. Then the analogy based effort estimation is done where the similar instances for a particular data is selected and from the similar instances, the mean value is calculated which gives the final effort. From the effort value calculated, the MRE value is found by finding the difference between the actual effort and the estimated effort. The mean value of the MRE gives the MMRE. In this research work, the Effort Estimation performance is validated based on MRE and MMRE, found out from Halstead Method, Bailey Method, Doty Method, COCOMO I and COCOMO II Models and the proposed model of K-Means clustering and analogy technique.

The number of clusters into which the dataset is to be divided is found using the knee point in the elbow plot. Figure 2 shows the number of clusters, k.

Table 1 shows the MRE values of the given dataset obtained by the methods such as Halstead Method, Bailey Method, Doty Method, COCOMO I and COCOMO II Models, Fuzzy method and the proposed model of K-Means clustering and Analogy based Estimation technique.

Table 2 shows the MMRE values of the given dataset obtained by the methods such as Halstead Method, Bailey Method, Doty Method, COCOMO I and COCOMO II Models, Fuzzy method and the proposed model of K-Means clustering and Analogy based Estimation technique.

Figure 3 shows the MRE values plotted for the methods such as Halstead Method, Bailey Method, Doty Method, COCOMO I and COCOMO II Models, Fuzzy method and the proposed model of K-Means clustering and Analogy based Estimation technique.

Figure 4 shows the MMRE values plotted for the methods such as Halstead Method, Bailey Method, Doty Method, COCOMO I and COCOMO II Models, Fuzzy method and the proposed model of K-Means clustering and Analogy based Estimation technique.

The MRE and the MMRE values must be as minimum as possible. By using the proposed method the effort can be estimated in the accurate manner and the estimated effort mostly coincides with the actual effort of the Dataset.

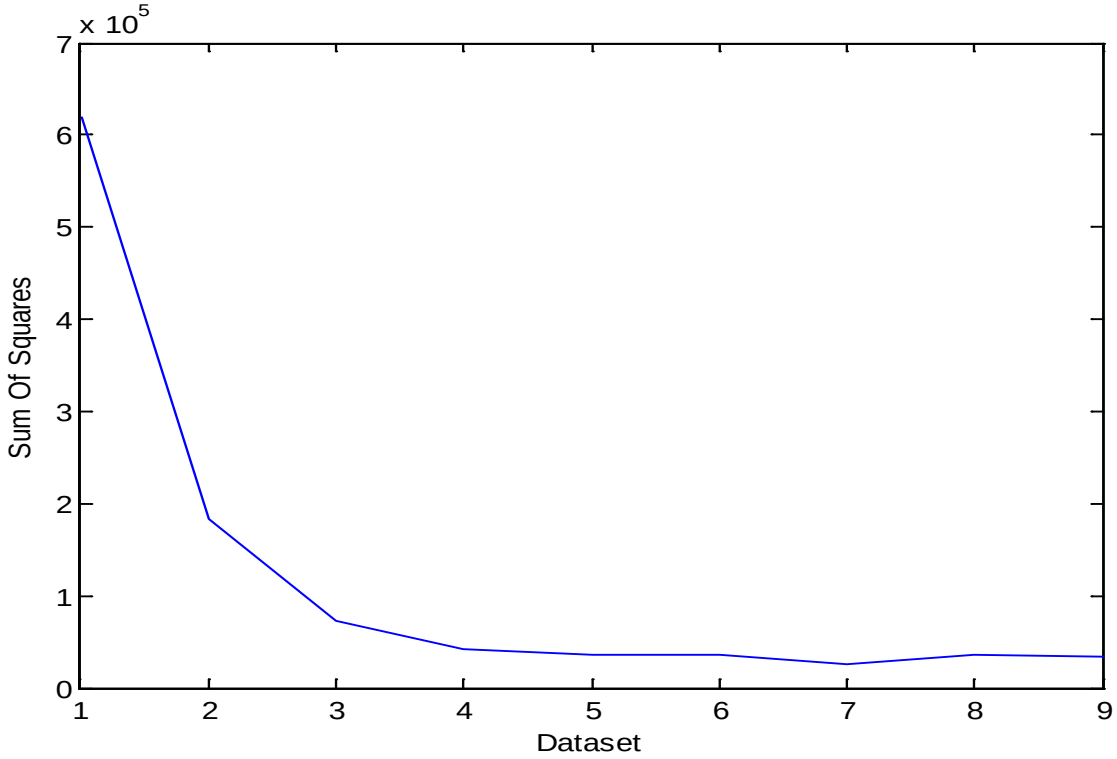

Figure 2: Elbow Plot for the PROMISE NASA Software Effort Estimation Dataset

Table 1: Comparing MRE Values

	MRE Values						
Dataset No.	**Halstead**	**Bailey**	**Doty**	**COCOMO I**	**COCOMO II**	**Fuzzy**	**Proposed Method**
1	1318.386	63.889	44.048	33.598	58.862	0.794	0.559807
2	2761.833	44.750	104.417	5.583	32.750	3.833	37.77551
3	11262.886	187.629	1044.330	427.835	321.649	1.031	145.1538
4	490.954	77.995	5.868	27.873	17.604	11.736	1.823077
5	639.687	77.383	6.543	25.747	16.074	6.970	3.032967
6	30.222	96.519	85.185	89.185	97.704	0.000	11.0989
7	318.125	83.542	29.583	46.458	72.292	0.000	3.7932
8	128.047	89.983	56.761	67.780	77.129	3.005	11.92517
9	42.326	91.860	65.581	75.349	82.326	4.419	62.33333
10	163.798	93.322	73.434	78.125	89.924	18.544	0.411538

Table 2: Comparing MMRE Values

MMRE Values						
Halstead	**Bailey**	**Doty**	**COCOMO I**	**COCOMO II**	**Fuzzy**	**Proposed Method**
1007.718	86.390	125.374	86.756	85.631	108.891	5.292

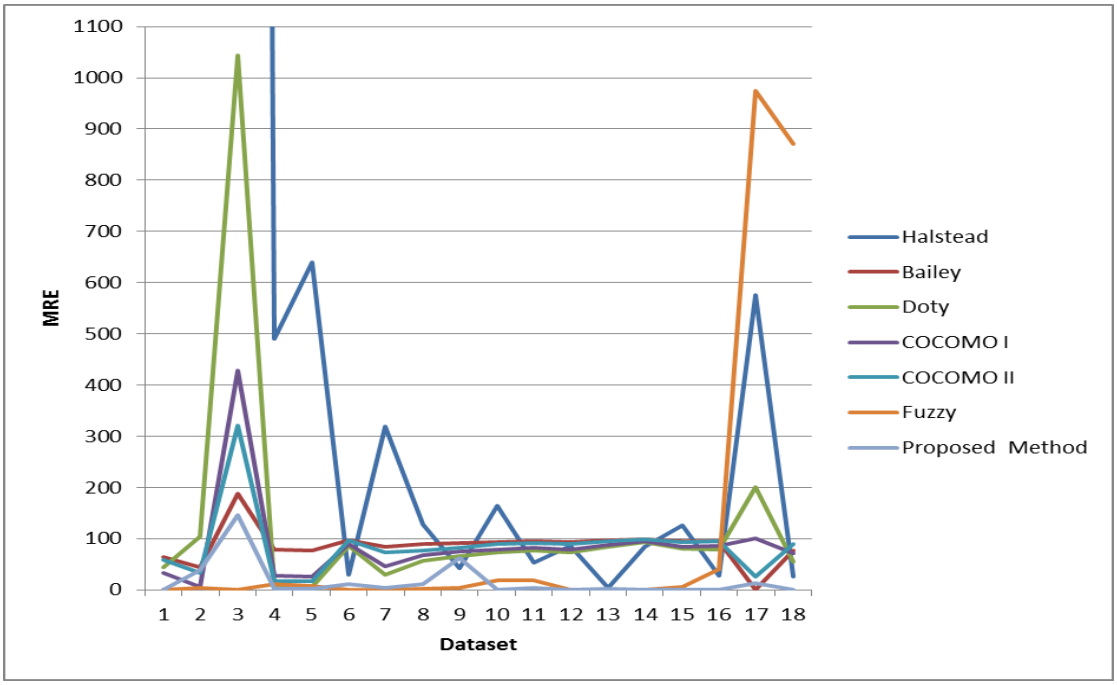

Figure 3: MRE Values Comparision

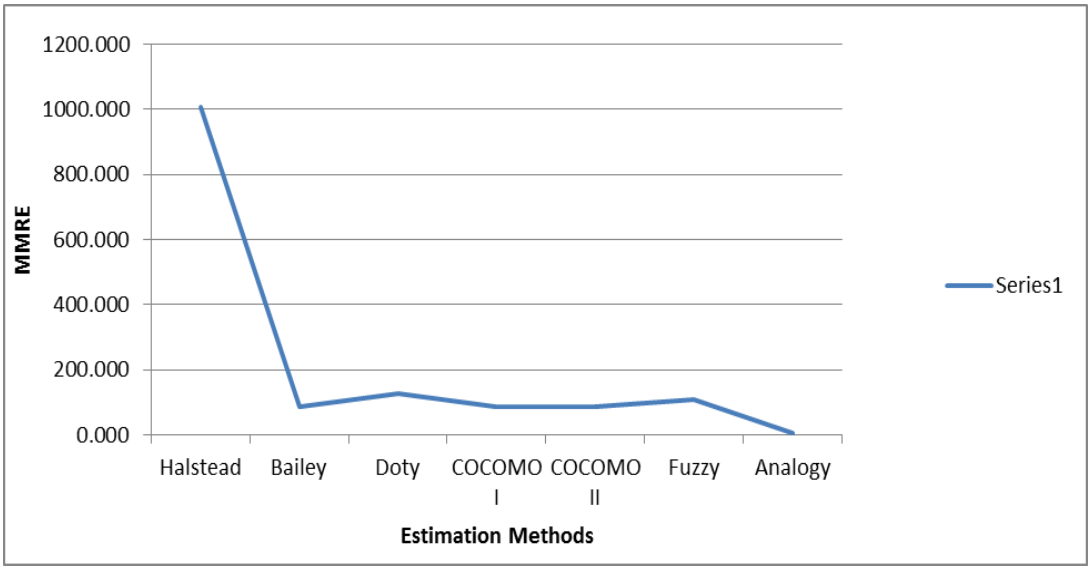

Figure 4: MMRE Values Comparision

VI. Conclusion

This project proposes an accurate and efficient way of estimating the effort. The results of the estimation based on the non-algorithmic method shows that the deviation between the actual and the estimated effort is very high. The result of non-algorithmic method of the analogy based estimation using K-Means Clustering reduces the Magnitude of relative error and the Mean magnitude of relative error. So the analysis of the effort from direct method and non-algorithmic method provides the inference that analogy based estimation using K-Means Clustering is the optimal method for the estimation of the software projects.

REFERENCES

[1] Khatibi E and Khatibi Bardsiri V, "Model to estimate the software development effort based on in-depth analysis of project attributes", Software, IET, vol.9 no.4, 2015, pp:109-118.

[2] Mrs.Nidhi Singh,Dr.Divakar Singh, "The Improved K-Means with Particle Swarm Optimization", Journal of Information Engineering and Applications, Vol 3,No11,2013.

[3] Mohammad Azzeh, Ali Bou Nassif, "Analogy-based effort estimation: a new method to discover set of analogies from dataset characteristics", IET Software,, Vol. 9, Iss. 2,2015.

[4] Jovan Zivadinovic et al., "Methods of Effort Estimation In Software Engineering", I International Symposium Engineering Management and Competitiveness 2011 (EMC2011), June 24-25, 2011.

[5] Rutvi Vanapalli,ch.Satyananda Reddy, "Predicting Error in Software Effort Estimate using K-Means and Quad Tree", International Journal of Computer Science and Information Technologies, Vol 5,2014.

[6] Ekrem Kocaguneli, Jacky Keung, "Active learning and effort estimation: Finding the essential content of software effort estimation data", IEEE Transactions on Software Engineering, vol.39, 2013, pp.1040-1053.

[7] Hari.CH.V.M.K,Tegjyot Singh Sethi,Kaushal.B.S.S,Jagadeesh.M, "SEEPC-A Toolbox for Software Effort Estimation using Soft Computing Techniques", International Journal of Computer Applications, Vol 31,2011.

[8] Youcheng Lin,Nan Tong,Majie Shi,Kedi Fan,Di Yuan,Lincong Qu,Qiang Fu, "K-means Optimization clustering Algorithm Based on Particle Swarm Optimization and Multiclass Merging", Advances in CSIE, Vol 1,2012.

[9] Jin-Cherng Lin,Yueh-Ting Lin,Han-Yuan Tzeng and Yan-Chin Wang, "Using Computing Intelligence Techniques To Estimate Software Effort", International Journal of Software Engineering and Applications, Vol 4,2013.

AUTHORS

First Author – M.Kowsalya, M.E(Computer Science and Engineering Department), Thiagarajar College Of Engineering, Madurai , kowsalya.mepco@gmail.com.
Second Author – H.OormilaDevi, M.E(Computer Science and Engineering Department), Thiagarajar College Of Engineering, Madurai , oormiladevi@gmail.com
Third Author – N.ShivaKumar, Assistant Professor(Computer Science and Engineering Department), Thiagarajar College Of Engineering, Madurai , shiva@tce.edu.