Pakistan Journal of
Life and Social Sciences

Enhancing the Software Effort Estimation using Outlier Elimination Methods for Agriculture in Pakistan

Nazish Murtaza, Ahsan Raza Sattar and Tasleem Mustafa
Department of Computer Science, University of Agriculture, Faisalabad-Pakistan

Abstract

Agriculture in Pakistan provides the means of living for more than 75 percent of its population. Any effort to improve crop production necessarily involves steps to manage the water management. It can be achieved using automated water management system. Many software are available for agriculture field but still there is lack of software that are used in precision agriculture; particularly in water management systems. As the progress increases in every field of life, the software project management has also improved. A poor estimate of effort and schedule is often suggested as a major contributor to software project failure. Most of the studies have also paid attention on the development of software effort estimation without consideration of outliers in data sets that cause the wrong results and decisions after implementing these software effort estimation methods. In this paper, we investigated the influence of outlier elimination upon the accuracy of software effort estimation through experiments applying two outlier elimination methods (K-means clustering and My-K-means clustering) and two effort estimation methods(Least squares and Neural network) associatively. This paper proposes a new outlier elimination method My-K-means clustering, which gives better estimation results than K-means clustering. The experiments are performed using the data of Agriculture in Pakistan, with or without outlier elimination. The estimated values of software effort showed the precision of research to improve the automated water management system. The experimental results are favorable because the minimum MMRE is 0.2078 and the maximum Pred (0.25) is 0.7454 using My-K-means clustering.

Keywords: Effort estimation, outlier elimination, water management introduction

Corresponding Author: Nazish Murtaza
Department of Computer Science, University of Agriculture, Faisalabad-Pakistan
Email: nazish.murtaza6@gmail.com

Introduction

Pakistan's crop sector consists of 15 to 16 major crops, which have been occupying almost 85 percent of the total cropped area for the last 25-30 years. The remaining 15 percent area is allocated to numerous other minor crops, vegetables and orchards etc (Shahnaz and Anwar, 2006). It is fact that the formers do not have much knowledge about water management in agriculture in Pakistan leading to inappropriate watering of crops. Sometimes excess of water causes the water logging and sometimes lack of water causes dryness of crops. So automated software can be developed to manage the water supply to the crops. Software engineering society has always faced the problems of accuracy of Software effort estimation.

Estimation accuracy and data quality correlates because without studying data quality, there is no impact of estimation accuracy. Most of the software effort estimation methods are built on the previous data (Jorgensen, 1995). However, these project history data contain outliers which can degrade project data quality. *Outlier* is unusual data value. The effects of outliers are bias or distortion of estimates and faulty conclusions.

In many cases, the software developers do not pay much attention or may not capable to estimate effort and time which is needed to develop the software accurately. Much software has ambiguous characteristics and unusual values which makes the estimates difficult. In this paper, we examined the influence of outlier elimination upon the accuracy of software effort estimation applying two outlier elimination methods (K-means clustering and My-K-means clustering) and two effort estimation methods (Least squares and Neural network) associatively. In our research, we estimated effort of software for water management in agriculture and used the data for experiment with or without outlier elimination.

The effort estimation methods applied in this paper are Least Squares and Neural Network. These methods have different theoretical background like statistics and machine learning respectively.

Least Squares refer to regression modeling in which the data is fitted to the pre-specified model. The overall sum of squared errors is minimized using the least square equation. Least squares method uses sample data to estimate the actual population relationship between two variables. This method is the most commonly used method for developing software estimation models (Jorgensen, 1995).

Neural network is an information processing technique based on machine learning. The network consists of input layers, hidden layers and output layers. NN is initialized with random weights for its arcs. These weights are used to reduce the difference between the predicted output and the actual output by training them gradually (Heaton, 2005). In this paper, we use feed-forward back-propagation type for software effort estimation using NN.

The outlier elimination methods applied in this paper are K-means clustering and My-K-means clustering. These methods have background of data mining.

A cluster is a collection of data that are similar to one another within the same cluster and are dissimilar to the data in other clusters (Han and Kamber, 2006). K-means sets K initial center points and partitions repeatedly the data into K mutually exclusive clusters until all clusters have the minimum total Euclidean distance between the data and their center point.

My-K-means clustering is an alternative method of portioning data. It requires more computation than K-means clustering but not required the number of clusters (k) and also gives better results than the K-means clustering. The goal of My-K-means clustering is to partition the large data into clusters and insure that each cluster has minimum possible number of data points.

In K-means clustering, the value of k must be given. So it produces different results when run each time. While in My-K-means clustering, it does not require specifying the number of clusters (k) and it always returns the same result when run several time.

The souvenir of this paper is organized as follows: Section 2 describes the related work. In Section 3, the overall approach of our experiment is presented. The final experimental result and discussion is explained in Section 4 and conclusion in Section 5.

Materials and Methods

The overall process consists of four steps shown in Figure 1.

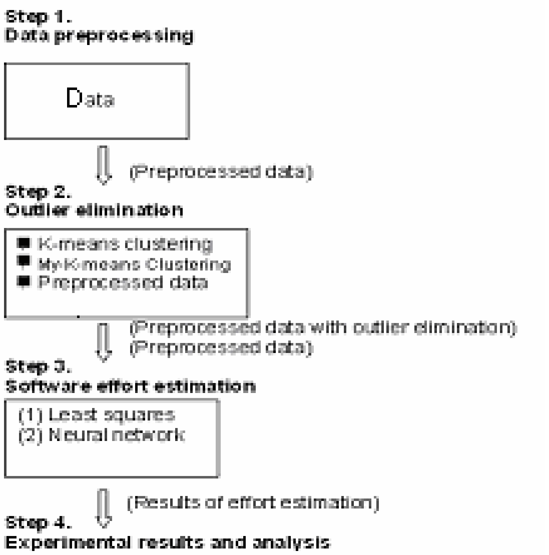
Explanation of target data set and experimental environment

The agriculture data set is used to our work to examine the effect of outlier elimination to the accuracy of software effort estimation. Statistical significance was set at 0.05.

Data preprocessing

The data was preprocessed to obtain the appropriate data set for the purpose of experiment by performing the selection of attributes and data elements from the data. 90 projects and 6 attributes were selected from the data set as shown in Table 1.

Figure 1. Overall process of our experiment



The data was excluded having zero or negative value on the attribute "Effort". After the selection the attributes, missing data handling was performed using imputation method using K-Nearest Neighbor (K-NN). After using the imputation method, the normality and correlation methods were applied to know whether the data set is under the assumption of statistical method because least square method was used as software effort estimation method. So the Shapiro-Wilk test was used to confirm the normality of Ratio-scaled attributes after log transformation of attributes. Then to find the correlation between the attributes, two-tailed Pearson's correlation test and One-Way ANOVA test was used.

Outlier elimination

In this step, two outlier elimination methods (K-means clustering and My-K-means clustering) are applied to the preprocessed data. The 12-fold cross validation was applied to data sets to obtain reliable results. The approach divides the whole data set into 12 folds and then 10 folds were used for the training sets and remaining two folds were used for the testing set.

After building the effort estimation model using 10 folds, compute estimation accuracy by the evaluation criteria (Section 3.5) using the remaining two testing folds. Finally, the accuracy results across all the folds are aggregated by average. Outlier elimination methods were applied on the data set. After outlier elimination, the silhouette value is calculated which ranges from -1 to +1. The data point which has the silhouette value less than zero and which is included in the cluster whose size is one, two or three was identified as outlier and then eliminated. In our work, the minimum cluster size is 4. Table 2 shows the number of outlier detected by K-means and My-K-means clustering on the data set.

Table 1. Variable description of the data

Name of Variable		**Description**	**Type**
Effort		Total project effort in person hours	Dependent
KAELOC		Total thousand assembly equivalent lines of code	Continues Independent
Max. Team Size (MTS)		Maximum number of members	
Project elapsed time (Duration)		Duration for the project in calendar days	
DP	**Development Platform Host (DP_Host)**	Dummy variable where “Host” platform is coded as 1 and others are coded as 0	Categorical Independent
	Development Platform Unix (DP_Unix)	Dummy variable where “Unix” platform is coded as 1 and others are coded as 0	
Model	**Life Cycle Model Inc (Model_Inc)**	Dummy variable where “Incremental” model is coded as 1 and others are coded as 0	
	Life Cycle Model V (Model_V)	Dummy variable where “V” model is coded as 1 and others are coded as 0	

Table 2. The number of detected outlier in Data

	1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold	Average
K-means	3	8	6	5	9	6	4	8	5	7	6.1
My-K-means	4	7	8	4	8	8	7	7	5	8	6.6

Software effort estimation

The 12-fold cross validation was applied to data sets to obtain reliable results. The approach divides the whole data set into 12 folds and then 10 folds were used for the training sets and remaining two folds were used for the testing set. Average value of the effort estimation accuracies using each testing set is used as the accuracy of the effort estimation model depending on each outlier elimination method and preprocessed data. The best fitting software effort models using LS presented as the following equation (Seo *et al.*, 2007):

$$\log(Effort) = 0.9222 + 0.1725 \square \log(KAELOC) + 0.5314 \square \log(Duration) + 0.6823 \square \log(MTS) + 0.0398 \square DP\ Host + 0.0571 \square DP\ Unix + (-0.0093) \square Model\ V + (-0.0424) \square Model\ Inc$$

In NN, all steps to build effort estimation model operate as black box, thus so we only know the results of effort estimation.

Experimental analysis

The effort estimation accuracy was measured using MMRE, MdMRE, Pred(0.25) and Pred (0.5) (Jorgensen, 1995). MMRE is one of the most widely used criteria. The estimation accuracy of MMRE≤0.25 and Pred(0.25)≥0.75 can be considered as acceptable levels of estimation accuracy (Conte *et al.*, 1986).

Results

The results show the accuracy of the effort estimation methods is increased by using the outlier elimination methods (Table 3). The MMRE is decreased in both the cases (Least Square and Neural Network) as well as Pred (0.25) is increased using outlier elimination methods. The experimental results are favorable because the minimum MMRE is 0.2078 and the maximum Pred (0.25) is 0.7454 using My-K-means clustering. The results also show that My-K-means clustering gives better results than K-means clustering. The accuracies of effort estimation models with outlier elimination are improved. The LS results on the data also present a small improvement of effort estimation accuracy while NN presents the good estimation accuracy results for outlier elimination method in terms of all evaluation criteria.

Table 3: Results for effort estimation methods with outlier elimination methods on the data

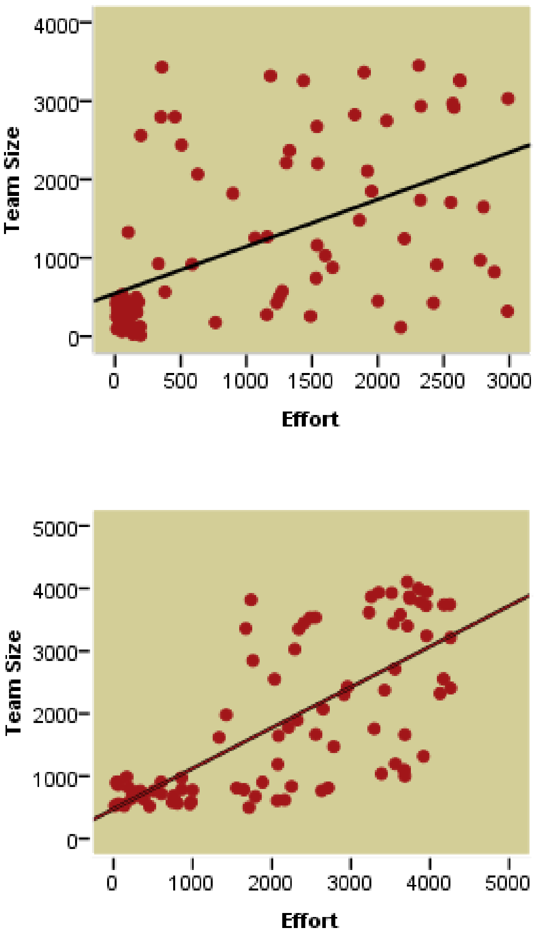
		Preprocessed	K-means	My-K-means
		Average from 10 folds		
LS	MMRE	0.3169	0.3067	0.2937
	MdMRE	0.2321	0.2152	0.2024
	Pred(0.25)	0.5436	0.5817	0.6279
	Pred(0.5)	0.8523	0.8454	0.8123
NN	MMRE	0.3156	0.2456	0.2078
	MdMRE	0.1859	0.1682	0.1482
	Pred(0.25)	0.6458	0.7169	0.7454
	Pred(0.5)	0.8512	0.8456	0.8238

Discussion

Our experimental results show that My-K-means clustering is more effective outlier elimination method than K-means clustering and it gives better estimation accuracy. In our research, it is noticeable that LS with outlier elimination methods is not significantly more accurate than LS without outlier elimination. NN provides the good estimation accuracy on the data. However, the outlier elimination methods applied to NN for the best estimation accuracy is different. The best choice is NN with the My-K-means on the data. Neural network performs well when the training set contains similar or redundant data point. This characteristic may have a great influence to the estimation accuracy of Neural Network.

A scatter pot is shown in Figure 2. It represents the quantitative relation between team size and development effort using linear regression model. Figure 2 shows that the accuracy of effort estimation is improved after using the outlier elimination.

Figure 2. Variation of effort estimation by outlier elimination



Conclusion

Effort estimation is done to improve the water management system for crop production in this research. The software effort estimation methods, which are built using the data samples with outliers, degrade the accuracy of effort estimation for software projects. Therefore, in this paper, a new outlier elimination method is proposed. We also examined the estimation accuracy of effort estimation methods when applying outlier elimination methods on data set. We used two outlier elimination methods and two effort estimation methods which have different theoretic backgrounds. Our study shows that the applied outlier elimination methods improve the estimation accuracy of software effort estimation models. In contrast, the effects of outlier elimination to the accuracy of effort estimation are different depending on the characteristics of data set, effort estimation methods and outlier elimination methods.

In our result, the My-K-means clustering gives better results. We also conclude that the application of Neural Network and My-K-means on the data set as the effort estimation method and outlier elimination method respectively present the most accurate software estimation results. It is also concluded that using the automated software for water management in agriculture, the production of crops increases.

References

Conte, S., Dunsmore, H. and Shen, V. Software Eng. Metrics and Models. Benjamin/Cummings publishing Company. 1986.

Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann. 2006.

Heaton, J. Introduction to Neural Networks with Java. Heaton Research, Inc. 2005.

Jorgensen, M. Experience with the accuracy of software maintenance task effort prediction models. IEEE Transactions on Software Engineering, 1995. 21: 674–681.

Seo, Y., Yoon, K. and Bae. D. The empirical analysis of Software Effort Estimation with Outlier Elimination. Software Engineering Laboratory, Division of Computer Science. South Korea. 2007.

Shahnaz, A. and Anwar F. Pakistan's Crop Sector: An Economic Evaluation. NWFP Agricultural University, Peshawar, Pakistan. 2006.