Journal of Scientific Research and Development 2 (4): 250-255, 2015
Available online at www.jsrad.org
ISSN 1115-7569

A hybrid approach artificial bee colony optimization and k-means clustering for software cost estimation

Ziba Ayat, Amin Babazadeh Sangar *

Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran

Abstract: Today, by increasing development of technology, different organizations and institutions require new software systems with proper quality. Along with this demand the importance of software systems increases everyday such that today software systems are among the most expensive computer systems. Therefore in order to address the requirements of different organizations software developing companies need to develop systems according to the requirements of their users. A project concludes successfully when it can address the financial requirements and demanded situations according to the contract between the two parties. For this, developing companies need to present a proper approach to estimate the primary cost. Primary approaches presented to solve this issue are mathematical methods among which central COCOMO which was developed by Bohem has a significant importance among the researchers. But noting that the presented methods were not able to solve the problems by updating the programming methods, this issue challenges the scientists that by proposing machine learning methods many researchers have presented new methods for machine learning but the presented methods were not able to estimate the price precisely. In order to increase the accuracy of cost estimation in software projects, in this work we tried to estimate software project costs by combining artificial honey bee colony and k-means clustering algorithms in a new algorithm and evaluated the results obtained from the combined algorithm compared to those obtained from central COCOMO model, artificial honey bee colony algorithm and k-means clustering algorithm based on the mean absolute relative error which showed increased precision in cost estimation.

Key words: Software cost estimation; Artificial honey bee colony algorithm; K-means clustering algorithm; Central COCOMO model

1. Introduction

By technological advances and in return increasing requirement to software systems in everyday life, development or extension of software systems has gained more importance. In order to develop a software system the project manager should be able to determine human sources (analyzer number, programmer, designer and supplier), equipment (required hardware and software), and time required for the completion of the project and ultimately according to obtained information estimate the cost in order to prevent the project from failure. Investigations have shown that by increasing the number of human sources and equipment, time period required for development and expansion of a system is decreased. Therefore the project manager should be able to make a balance between these three factors, and according to the time determined by the client, he should manage the human sources and equipment in the best possible way. After determining the time, human sources and equipment, the project manager should estimate the cost of producing such a system according to these factors. It should be kept in mind that precise estimation of the cost can determine the success or failure of the system. In fact the accuracy of the estimated cost is highly significant in efficiency of the organization (Gharehchopogh and Pourali, 2015; Gharehchopogh and Maroufi, 2014; Maleki et al., 2014). Therefore because the cost estimation is performed at the beginning of the project and at that time there is not enough information available about the project, the cost cannot be estimated fully precisely. Therefore this issue appears as a problem in software engineering. For this reason many researchers have presented ideas to solve this problem. Generally the solution can be expressed as follows:

1. The expert: the expert is among the most successful solutions. But since in this method we need a highly experienced person and extensive knowledge and such a person is not available for all organizations, this is a challenging solution.
2. Mathematical and formulated solutions: solution of software cost estimations using mathematical formulas began in 1970 and the most documented and addressed method in this group is COCOMO 2 model. COCOMO model was presented by Boehm in 1981 (Boehm, 1981). The formula concerned in COCOMO is consisted of three constant parameters, factors effective in software cost

* Corresponding Author.

estimations and project size. Bu according to the increasing complexity and the volume of software projects and appearance of new components in software development and extension, solution of cost estimation by mathematical procedures is not suiTable.

3. Machine learning methods: in machine learning methods researchers use data analysis algorithms, Evolution algorithms, etc. to solve different problems. In these methods the main objective is to write a program capable of improving its performance by learnt experience. Learning may resultin change in program and/or data structure (Karaboga and Akay, 2009).

The remaining of the article is organized as follows. Section 2 reviews previous works. Section 3 presents basic concepts (middle COCOMO model, artificial bee colony algorithm, clustering algorithm). The proposed method is presented in section 4. Section 5 evaluates the performance of the proposed model and finally conclusion and future works are presented in section 6.

2. Previous works

Today projects related to official software systems have gained significant importance, therefore one should be careful that is the required time and cost for the development of software system is not estimated precisely possibility of financial loss and project failure is increased. According to this software project cons estimation is one of the most important factors in development and extension of software systems. Such that since late 1970 the researchers have tried to solve this problem by presenting different methods some of which can be mentioned as follows**?**

In (Tegjyot Singh Sethi, 2011) in order to estimate the cost of software projects PSO and k-mean clustering algorithms have been used. These researchers used k-mean clustering algorithm in clustering rather than manual classification of datasets. In fact the values of constant parameters of COCOMO 2 model which were predicted by PSO algorithm were different for each cluster and since the datasets were clustered it was expected that the produced results also be better. In this study datasets used in COCOMO 81 were compared to the results of standard COCOMO 2 which showed better performance of the method**.**

Dizaji and Gharehchopogh (2015) used hybrid of Ant Colony Optimization and Chaos Optimization for Software cost estimation. In this article, ant mapping was used as the chaos factor and NASA datasets were employed. In each step of the optimization each of the algorithms were executed separately and then the algorithm with the best result was used as the solution for cost estimation in that step. The results obtained from the presented method were compared to those of COCOMO model which is known as the most documented mathematical model in project software cost estimation. It is noteworthy that the evaluation criterion in this article is mean absolute relative error. According to the obtained results the performance of PSO algorithm was better than that of COCOMO model and the performance of combined algorithm was better than both COCOMO model and PSO algorithm **.**

F.S.Gharehchopogh (2011) used artificial neural network in software cost estimation. In this reference 11 projects of 60 available projects in NASA dataset were compared to COCOMO model and the results showed that in the majority of cases the error in COCOMO algorithmic model was more than ANNs models. According to the results in more than 90% of the cases the presented method made better estimations compared to COCOMO algorithmic model for the projects available in NASA dataset. Therefore it could be concluded that methods based on artificial intelligence are good alternatives to algorithmic methods**.**

In (DIZAJI et al., 2014) researchers used artificial bee colony optimization algorithm in software cost estimations. These researchers used artificial bee colony optimization algorithm to assign values for constant parameters a and b and after assigning values to these parameters used them in software cost estimation based on COCOMO model. In order to evaluate the accuracy of the estimated cost these researchers evaluated the presented model based on mean absolute relative error and the results showed increases estimated costs.

3. Basic concepts

In this section we first discuss software cost estimation and then artificial bee colony optimization algorithm and finally k-means clustering algorithm.

3.1. Software cost estimation and middle COCOMO model

In the past four decades, many quantitative models have been developed in order to evaluate software costs. Generally these models can be classified in two classes; first class being algorithmic models and the second being non-algorithmic models. In the first class mathematical formulas have been used in estimations. These formulas have different complexities such that they can be very simple such as statistical summary with modest deviation (Donelson, 1976) or very complicated such as models based on regression (Walston and Felix, 1977) or models based on differential (Putnam, 1978). In this model in order to increase the accuracy, the information should be adjusted and calibrated according to local situations. In the second class, unlike the first one, analytical and inferential comparison has been used in cost estimation**.**

The most documented method among algorithmic models is COCOMO model which was first introduced by Boehm in 1981 (Boehm, 1981). COCOMO model has three different types: primary COCOMO model, middle COCOMO model, and advanced COCOMO model among which the majority

of the researchers prefer middle COCOMO model in comparing with their own methods and the equation used in calculating cost estimation is as (1) (Menzies et al., 2005):

(1): $PM = a * (size)^b * \prod_{i=1}^{15} EM_i$

Parameters a and b have constant values and are determined according to the tee of the project.

3.2. Artificial bee colony algorithm

Artificial bee colony is considered as a collective intelligence algorithms. This algorithm was first introduced by Karabage to optimize mathematical functions. In ABC algorithm each food source determines a solution (Panigrahi et al., 2011) whose syrup content shows the competence of the presented solution. In this algorithm bees are classified in three classes: worker bees, searcher bees, and pioneer bees**.**

Searcher bees chose a food source according to its corresponding probability. In case the syrup content of the food source is high its probability to be chosen is higher. This probability can be calculated using equation (2):

(2): $p_i = \frac{fit_i}{\sum_{j=1}^{n} fit_j}$

In equation (2) variable $f_i$ti determines the competence value of the i^{th} solution; the competence value of each solution is evaluated by its worker bee. In fact this evaluation is based on the syrup content of the food source in i^{th} location and variable N shows the number of food sources. In this method in order to search a chosen location of a food source, worker bees exchange their information with searcher bees and to do these exchanges equation (3) is used (N.Ebrahimpour et al., 2015), (DIZAJI et al., 2014):

(3): $V_{i,j} = X_{i,j} + \varnothing_{i,j} * (X_{i,j} - X_{k,j})$

In equation (3), k, iϵ{1,2,...,EB } and jϵ{1,2,...,D} are randomly chosen and variable øij is a random value in [-1, 1] which controls the location of the food source produced in neighbor Xij.

3.3. K-Means clustering algorithm

Clustering is done in many methods and one of the most commonly used clustering methods is k-means clustering method. This clustering method was presented in 1967 by Mc Queen (MacQueen, 1967). In this clustering method it is tried to make two estimations which can be expressed as follows:

Determination of clustering centers that are in fact the mean of points related to the clusters.

Assigning any sample data to the cluster such that the highest similarity with the remaining members in the cluster is achieved.

In this algorithm equation (4) is used to determine the similarity content of one sample data to the cluster (Zhang and Fang, 2013, Gharehchopogh and Dizaji, 2014):

(4): $\|X_i - Z_j\| < \|X_i - Z_p\|$,

$P = 1,2,\dots,K$ and $j \neq p$

In equation (4) variable Xi shows the sample data, Zj shows the investigated cluster center and Zp shows other clusters.

After temporary clustering sample data, the data cluster centers are recalculated with equation (5).

(5): $Z_i = {}^{1}/_{N_i}$

In equation (5) variable Ni shows the number of samples existed in cluster Zi.

4. Proposed method

In this paper the combination of k-means clustering and artificial bee algorithm have been presented in estimating software project costs. In the presented combined method first the NASA datasets were divided to test and control datasets; 80% of the data we assigned as learning data and 20% was assigned to be test data and k-means clustering algorithm was used to clustering the projects.

Once the data was concluded, artificial ant colony algorithm was used to assign values to constant variables in COCOMO model and in order to evaluate the performance of each bee, the criterion mean absolute relative error was used as the fitness function and in each step of performing the algorithm it was cried to reduce the value of this variable. After learning step, the values of the constant variables of middle COCOMO model were assigned for different clusters. After assigning values to these parameters, costs of the existing projects in the test dataset were estimated by parameters produced for each cluster. The performance of the presented method is shown in Fig. 1 and Table 1.

Table 1: Performance of combined algorithm

1. Input:
NASA dataset and parameters related to artificial bee algorithm and k-means clustering algorithm
2. Output:
Constant parameters related to middle COCOMO model and estimated costs for software projects
3. Steps:
First step: reading the data in NASA dataset
Step two: dividing datasets into test and learning data
Step three: clustering test and learning data using k-means clustering algorithm
Step four: determination of the number of food sources, worker bees, searcher bees, and pioneer bees
Step five: resenting primary solutions by pioneer bees
Step six: determining the quality of the solutions presented to the searcher bees
Step seven: calculation of probability of choosing the solutions
Step eight: choosing the solutions greedily
Step nine: optimizing the presented solutions
Step ten: investigating the solutions by searcher bees
Step eleven: in case the solution is abandoned, a new solution is produced by the pioneer bee

Step twelve: is it time to stop the algorithm?
Step thirteen: if not go to step six
Step fourteen: presenting the values for the constant parameters (a, b)
Step fifteen: cost estimation for the existing projects in test dataset according to the constant parameters
Step sixteen: evaluation of estimated costs with mean absolute relative error

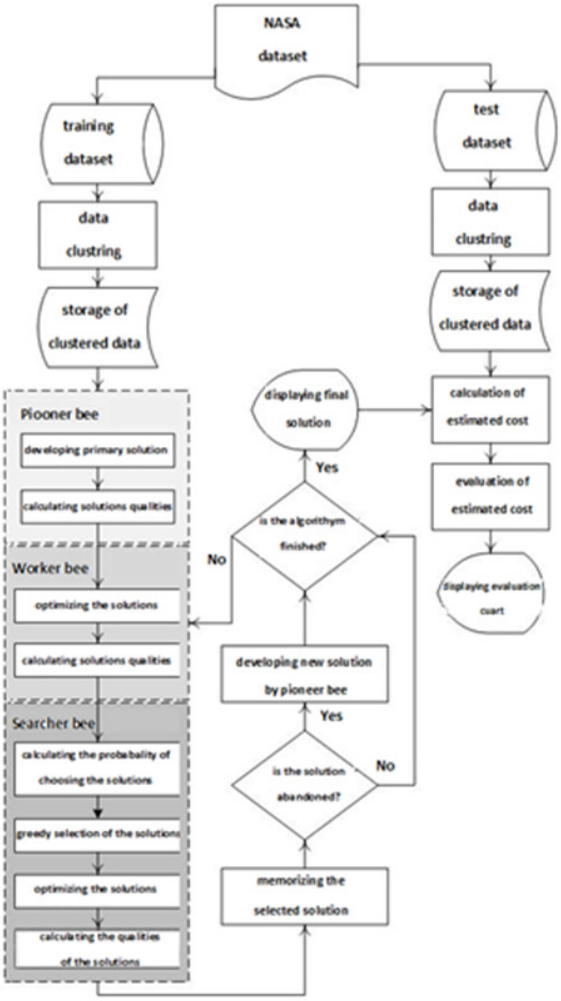

Fig. 1: Performance of combined algorithm

5. Results and discussion

Precise cost estimation of software projects is a determining step in failure or success of a software project, therefore in this article in order to increase the accuracy of the cost estimation for software projects, a combined method using artificial bee algorithm and k-means clustering algorithm was used on NASA datasets and the obtained results also showed that the cost estimation accuracy was improved. It is noteworthy that in this article the evaluation was done according to the criterion mean absolute relative error which is calculable using equations (6) and (7).

(6): $\mathrm{MARE_i} = \frac{|\mathrm{Actual_i} - \mathrm{Estimate_i}|}{\mathrm{Actual_i}}$

(7): $\mathrm{MMARE} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{MARE_I}$

In Fig. 2 in order to evaluate the proposed method, this method was compared to middle COCOMO model which was done on learning datasets and its improved performance can be inferred from Fig. 2.

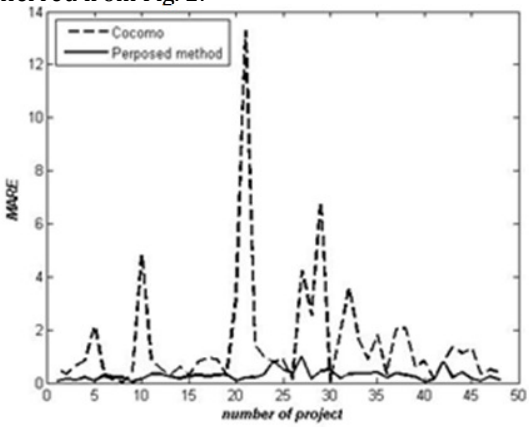

Fig. 2: Comparison of the performance of combined algorithm and COCOMO model on train data

Fig. 3 compares the performance of the proposed method with that of middle COCOMO model which was performed using test datasets. According to the obtained results it can be concluded that the proposed method has performed better.

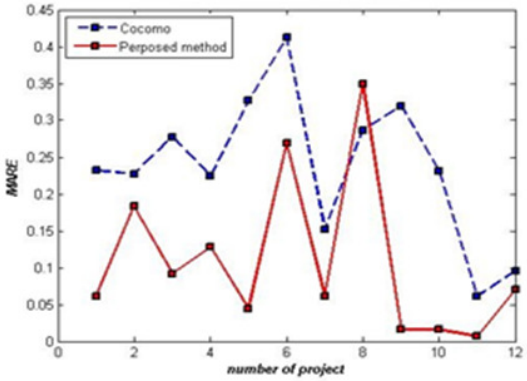

Fig. 3: Comparison of the performance of the combined algorithm with that of COCOMO model on test datasets

Fig. 4 shows the performance of artificial bee algorithm, the proposed method and COCOMO model. According to these results it can be said that the estimated cost using the proposed method had more accuracy for the majority of the software projects compared to the other algorithms.

Table 2 shows the mean MMARE error for artificial bee algorithm, the proposed method and COCOMO model. The value of this error for the proposed method is minimum compared to the other methods which shows the maximum accuracy in cost estimation.

Table 2: Evaluation according to criterion MMARE error

Model	MMARE error
COCOMO model	0.2371
Artificial bee	0.1380
Proposed algorithm	0.1081

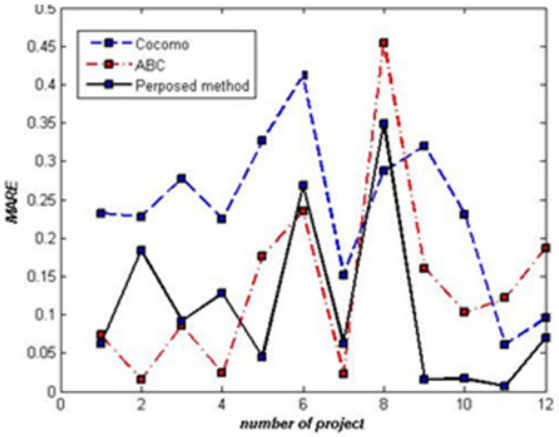

Fig. 4: Comparison of the considered algorithms on test datasets

6. Conclusion and Future Works

Each year much software are developed and expanded according to the requirements of the users in the world. Meanwhile software producing organizations need to use proper methods in estimating the primary costs for the software to be produced in order for them to increase the profitability. But so far a method capable of providing fully precise estimation without any error has not been introduced. In this article in order to reduce the error in cost estimation we have presented a new method based on machine learning in which two artificial bee colony and k-means clustering algorithms have been combined and the results obtained from the proposed method have been evaluated using criterion mean absolute relative error. In the combined algorithm the projects have been classified with k-means clustering algorithm rather than the classification according to the type of the projects. According to the obtained results it can be said that artificial bee colony increased the accuracy of the cost estimation. But when this algorithm was combined with k0-means clustering algorithm the accuracy of the estimated cost became even better such that it can be said that the proposed method resulted in 13% more accurate cost estimation compared to middle COCOMO model.

References

B.W.Boehm (1981), "Software Engineering Economics", Prentice- Hall, Englewood Cliffs, New Jersy.

Boehm. B. W (1981), "Software Engineering Economics", Prentice- Hall.

C. E. Walston and C. P. Felix (1977), "A method of programming measurement and estimation", IBM Systems Journal, vol. 16, no. 1, pp. 54-73.

C.Zhang, Z.Fang (2013),"An Improved K-means Clustering Algorithm", Journal of Information & Computational Science, vol.10: 1, pp. 193–199.

D.Karaboga, B.Akay (2009), "A comparative study of Artificial Bee Colony algorithm", Applied Mathematics and Computation 214, vol, PP.108–132.

F.S. Gharehchopogh (2011), "Neural Networks Application in Software Cost Estimation: A Case Study", International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2011), pp.69-73, IEEE, Istanbul, Turkey.

F.S. Gharehchopogh, A. Maroufi, (2014) "Approach of software cost estimation with hybrid of imperialist competitive and artificial neural network algorithms", JOURNAL OF SCIENTIFIC RESEARCH AND DEVELOPMENT, Vol: 1, No: 1, pp. 50-57.

F.S. Gharehchopogh, A. Pourali,(2014 "A New Approach Based on Continuous Genetic Algorithm in Software Cost Estimation", JOURNAL OF SCIENTIFIC RESEARCH AND DEVELOPMENT, Vol: 2, No: 4, pp. 87-94, 2015.

F.S. Gharehchopogh, Z. A. Dizaji, (2014), "A New Approach in Software Cost Estimation with Hybrid of Bee Colony and Chaos Optimizations Algorithms", MAGNT RESEARCH REPORT, Volume 2, Issue 6, pp: 1263-1271, Nov 2014.

I. Maleki, F.S. Gharehchopogh, L. Ebrahimi, Z. Ayat, (2014) "A Novel Hybrid Model of Scatter Search and Genetic Algorithms for Software Cost Estimation", MAGNT RESEARCH REPORT, Volume 2, Issue 6, pp: 359-371, 2014.

J. B. MacQueen (1967), "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.

L. H. Putnam (1978), "A general empirical solution to the macro software sizing and estimating problem", IEEE Trans. Soft. Eng, pp. 345-361.

Menzies.T, Port.D, Chen.Zh, Hihn.J (2005), "Validation Methods for Calibrating Software Effort Models", ICSE ACM USA.

N. Ebrahimpour, F.S. Gharehchopogh, Z.A. Khalifehlou, "A New Approach with Hybrid of Artificial Neural Network and Ant Colony Optimization in Software Cost Estimation", MAGNT RESEARCH REPORT, Vol 3, No: 2, pp: 1081-1089, Feb 2015.

Panigrahi, K.Bijaya, Sh.Yuhui, L .Meng-Hiot (2011), "andbook of Swarm Intelligence ", Springer-Verlag Berlin Heidelberg.

Tegjyot Singh Sethi (2011), CH.V.M.K.Hari , B.S.S.Kaushal , and Abhishek Sharma ," Cluster

Analysis and Pso for Software Cost Estimation" , Communications in Computer and Information Science, vol 147 , pp 281–286.

Z. A. Dizaji, F.S. Gharehchopogh,(2015), "A Hybrid of Ant Colony Optimization and Chaos Optimization Algorithms Approach for Software Cost Estimation ", Indian Journal of Science and Technology, Vol: 8, No: 2, pp: 128-133.

Z. A. Dizaji, R. Ahmadi, H, Gholizadeh, F.S. Gharehchopogh (2014), "A Bee Colony Optimization Algorithm Approach for Software Cost Estimation" International Journal of Computer Applications (IJCA), Vol: 102.