Software Effort Estimation for NASA Projects Using Genetic Programming

Alaa F. Sheta[1], Alaa Al-Afeef[2]
[1]The World Islamic Sciences and Education University (WISE)
Amman, Jordan
alaa.sheta@wise.edu.jo
[2]Image Technologies Inc (ITEC)
Amman, Jordan
alaa.afeef@gmail.com

ABSTRACT*: There is still an urgent need of finding a mathematical model which can provide an accurate relationship between the software project effort/cost and the cost drivers. A powerful algorithm which can optimize such a relationship via developing a mathematical relationship between model variables is urgently needed. In this paper, we explore the use of Genetic Programming (GP) to develop a software cost estimation model utilizing the effect of both the developed line of code and the used methodology during the development. An application of estimating the effort for some NASA software projects is introduced. The performance of the developed GP based model was tested and compared to known models in the literature. The developed GP model was able to provide valuable estimation capabilities compared to other models.*

Keywords: Software Cost Estimation; Software Engineering; Genetic Programming; NASA Software

Received: 11 May 2010, Revised 18 June 2010, Accepted 25 June 2010

1. Introduction

Reliable predictions of project costs primarily effort are greatly needed for better planning of software projects.

The software project manager should be able to reliablely estimate the overall project costs, duration, required man power and schedule [1]. He must be able to fairly distribute the resources over time such that the project could be finished on time and within budget. It was found that there are many similarities between the process of managing project resources and system modeling. In system modeling we need to develop some sort of a relationship between the system input and output such that the system function is approximated in a form of a model. The model can be used for simulation and performance evaluation of the original system under various operating conditions. In project management, the manager need to collect enough data about various attributes which affect the quality and the cost of a project. These collected data helps in developing a plan or a model for cost distribution over various phases of a project. The developed model can be calibrated in each phase of the project to meet the project goals, the quality of the product and the available resources.

1.1 The estimation of software effort

Software effort estimation process has a similar nature since it is part of project management. In this case, the objective is to develop a sort of relationship between the expected Developed (DL) Line Of Code of a project as an input variable and the expected effort required to implement this project in man-month. There has been extensive research into software effort

estimation, with researchers assessing a number of approaches to improving prediction accuracy. One of the famous effort DL-E relationship [2], [3] known as the COnstructive COst MOdel (COCOMO) is give as in Equation 1.

$$E = a(DL)^b \quad (1)$$

The DL include all program instructions and formal statements [4]. The values of the parameters *a* and *b* depend mainly on the class of software project. Software projects were classified based on the complexity of the project into three categories. They are: 1) Organic 2) Semidetached and 3) Embedded. COCOMO model was first provided by Boehm [2], [5]. This model was built based on 63 software projects. The model helps is defining mathematical equations that identify the the cost, schedule and quality of a software product. The estimation accuracy is significantly improved when adopting models such as the Intermediate and Complex COCOMO models [2]. Extensions of COCOMO, such as COMCOMO II, can be found in [3].

Typical models for software effort estimation are given in Table I. These models have been derived by studying large number of completed software projects from various organizations and applications to explore how project sizes mapped into project effort.

Model name	Model equation
Halstead	$E = 5.2(DL)^{1.50}$
Walston-Felix	$E = 0.7(DL)^{0.91}$
Bailey-Basili	$E = 5.5 + 0.73(DL)^{1.16}$
Doty (for DL > 9)	$E = 5.288(DL)^{1.047}$

Table 1. Known Effort Estimation Models

1.2 Previous Work

In the past, most of the proposed models used to solve the software cost estimation modeling problem are linear in nature. It was found that dealing with a linear model makes it easier to use techniques such as least square estimation (LSE) or Instrumental Variable method to identify the parameters of the given model. In the other case, if the actual model is nonlinear, attempting to approximate this structure with a linear model cannot guarantee the accuracy of the model. In solving the software cost estimation problem, it is important to develop models using a small number of measurements and in the presence of measurement noise.

Recently, many questions were introduced about the applicability of using Soft Computing and Machine Learning Techniques to solve the effort and cost estimation problem for software systems. In [6], [7], authors presented a detailed study on using number of techniques such as genetic programming and neural networks to estimate software project effort. Author concluded that GP can perform well on handling such a problem. In [8], author provided an innovative set of models modified from the famous COCOMO model with interesting results. Later on, many authors explored the same idea with some modification [9]–[12] and provided a comparison to the work presented in [8]. In [13], author used Particle Swarm Optimization (PSO) to tune the parameters of the famous COnstructive COst MOdel (COCOMO). They also explored the advantages of Fuzzy Logic to build a set of linear models over the domain of possible software Line Of Code (LOC). The performance of the developed model was evaluated using NASA software projects data set. Also a variety of machine learning methods have been used such as case based reasoning (CBR) [14], [15], rule induction (RI) [16] and Hybrids [17].

In this paper, an evolutionary approach, Genetic Programming (GP), is used to fit nonlinear models to a dataset of some NASA software projects, aiming to improve the prediction of software effort for NASA software projects.

In this paper, Genetic Programming is used to develop an effort estimation model for software systems due to the advantages of GP as provided in Section II-A. The theoretical foundations of genetic programming are summarized in [18].

In the following Section II, GP is introduced briefly. The experiment setup and control parameters for the application of GP in evolution of software development effort estimation programs is discussed in Section III and the developed results in Section IV. This includes data preparation, GP details and results obtained. A comparison of related developed results are presented in Section V. Section VI draws the conclusions and future work.

2. Overview of Genetic Programming

Genetic programming (GP) is an evolutionary computation (EC) technique that automatically searches for an optimal solution of a problem without requiring the user to know or specify the form or structure of the solution in advance [19], [20]. GP technique has been successfully applied to solve large number of difficult problems, such as modeling of industrial processes [21], [22], forecasting of river flow [23], image reconstruction [24], [25] and Generating models to fit data [26]–[28].

2.1 Advantages of using Genetic Programming

Evolutionary algorithms have been found 'experimentally' efficient in finding solutions to the Modeling problems. GP is considered one of the evolutionary algorithms that hold all advantage offered by evolutionary algorithms and adds several more. The advantages offered by GP for Modeling can be summarized as:

- GP is a global search technique that makes use of hyper plane search which, makes it less likely to get stuck in the local optimum. This is different from other techniques such as neural networks and gradient descent which are prone to local optimal values.
- GP has the benefits of variety in solution structures unlike most of the evolutionary algorithms that has fixed size solutions such as genetic algorithms or fixed architectures such as neural networks [29].
- GP can automatically eliminate unrelated attributes of the Modeling problem performing the task of feature extraction algorithm [29] in which important attributes can appear near the root while less important ones would appear deeper in the tree [30].
- GP is able to operate on portion of data to extract significant rules. There is no need to use all of the training data to develop models [29].
- GP are like white boxes that clearly sketch the relationships between attributes, as opposed to many other black box solutions like neural networks [31].
- GP has the ability to operate upon the data in its original form. No pre-processing or data transformations are usually required to apply GP for modeling task.
- GP based evolution is not affected by the data distribution [29]. This is in contrast to the neural networks which are highly dependent on the data distribution. This autonomy enables efficient discovery of unknown knowledge from the data.

2.2 Representation in GP

In GP, programs are usually represented as a variable sized tree structure. This type of representation allows a variety of models to be developed. A tree consists of nodes and terminals. In every terminal node, there is an operand and in every node there is a function. Trees can be easily evaluated in a recursive manner. This way we can evolve mathematical models in a very simple way such as in programming using Lisp language [32]. Such a representation is simple and has been used frequently for the data classification and modeling problems. A simple tree structure can be presented in Figure 1 as described in Equation (2).

$$E = 1.7 \cdot DL - ME \qquad (2)$$

2.3. Preparatory Steps of GP

Before applying the Evolutionary Process, as in Figure (2), four major preparatory steps require to be specified [19], [20]:

1) The definition of the function and terminal set (primitive set) for a particular problem.

2) Fitness measure for the problem. This specifies what needs to be done.

3) The control parameters for the run (for example,population size, max generations and maximum tree depth).

4) The termination criterion which may include a maximum number of generations to be run as well as a problem-specific optimum solution.

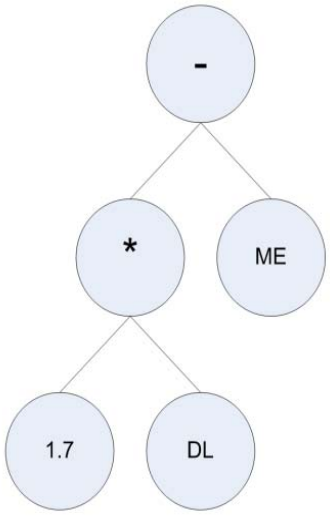

Figure 1. Example of GP tree structure

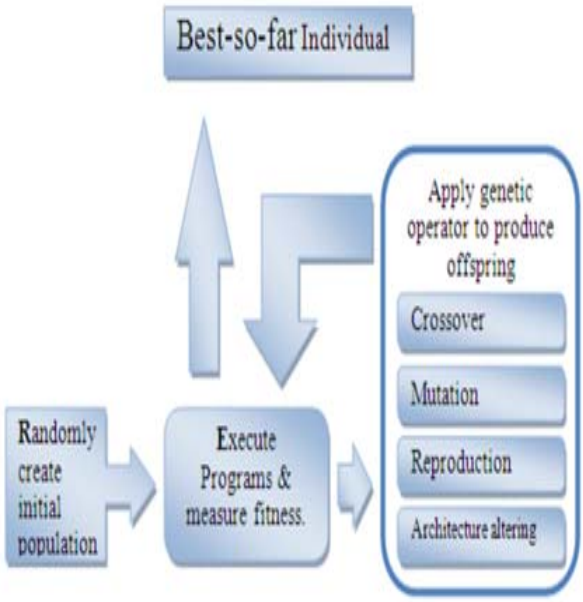

Figure 2. GP evolutionary process

2.4. Performance evaluation Criteria

In order to check the performance of the developed models, two evaluation criteria will be adopted. We compute the Variance-Accounted-For (VAF) performance criterion to measure how close the measured values to the values developed using the fuzzy models. Given that E, $\hat{E}$ are the actual effort and the estimated effort, respectively. The VAF is computed as follows:

$$VAF = [1 - \frac{var(E - \hat{E})}{var(E)}] \times 100\% \quad (3)$$

The Mean Magnitude of Relative Error (MMRE) as the main performance measure was also used in many articles [13], [33]. MMRE is defined as:

$$MMRE = \frac{1}{N} \sum_{i=1}^{N} \frac{|E - \hat{E}|}{|E|} \quad (4)$$

We will also adopt these two criteria's for evaluating the cost estimation models investigated here.

3. Experiment Setup and Control Parameters

GP Setup (Table II) is adapted for modeling the problem under study. The adopted control parameters are shown in Table IV and Table VI according to [19].

Experiments have been conducted on a data set presented by Bailey and Basili [34] to explore strengthen of the developed GP based model. The dataset consist of the following variables:

- Developed Line of Code (*DL*)
- Methodology (*ME*) and
- Effort (*E*) in man-month.

The dataset is presented in Table III. The data was split to two sets training (i.e. 13 projects) and testing/validation (i.e. 5 projects). We used Lilgp1.1 [35] (C language package for developing genetic programming applications) to produce our results. Lilgp is well-known to be a fast, memory efficient and well documented GP tool that provides support for several features not typically found in other GP systems, such as the support of parallel processing.

Objective	Find a function of 2 independent variable[Line Of Code (DL), Methodology(ME)] and one dependent variable [Effort (E)], in symbolicform, that fits a given Training sample of the form (DL, ME, E) data points.
Terminal set	DL, ME (the independent variables).
Function set	+, -, *
Fitness criteria	The fitness is the absolute value of the difference between the estimated values produced by GP and the target value of the effort. ($\lvert E^i_{Target} - E^i_{Estimated} \rvert$).
Raw fitness	The sum taken over the fitness cases (N) ($\sum_{i=1}^{N} E^i_{Target} - E^i_{Estimated}$)
Standardized fitness	Equals raw fitness divided by the count of fitness cases.
Hits	Number of fitness cases for which the valueof the dependent variable produced by theGP comes within 0.001 of the target value.

Table 2.GP Experiment setup for the Effort Estimation Problem

Project No.	DL	ME	Effort E
1	2.1	28	5.0
2	3.1	26	7.0
3	4.2	19	9.0
4	5.0	29	8.4
5	7.8	31	7.3
6	9.7	27	15.6
7	10.5	34	10.3
8	12.5	27	23.9
9	12.8	26	18.9
10	21.5	31	28.5
11	31.1	35	39.6
12	46.2	20	96.0
13	46.5	19	79.0
14	54.5	20	90.8
15	67.5	29	98.4
16	78.6	35	98.7
17	90.2	30	115.8
18	100.8	34	138.3

Table 3. Sorted Nasa Software Project Data

3.1 GP Effort Model based DL

The developed GP model should be able to significantly generalize the computation of the developed effort for all projects. We run GP to develop a new software effort estimation model. The developed Lisp expression program is given in Equation 5 which is simplified in Equation 6.

$$(*(-(+(1.35730DL)1.75992)(*1.36186DL))DL) \quad (5)$$

$$E = 1.75992 \cdot DL - 4.56 \cdot 10^{-3} DL^2 \quad (6)$$

We run GP with various population sizes (i.e. 1000, . . . ,9000). The convergence process for all runs were measured and the best so far curves are presented in Figure (3). It is shown that all curves convergence to the same optimal value for the fitness criteria. The rest of the tuning parameters for the Lilgp experimental setup is given in Table IV. Table (V) show the measured and estimated GP effort.

3.2 GP Effort Model based DL and ME

GP was used to find the model structure which describe the relationship between the effort and both the developed line of code and the methodology. We run GP was various population sizes to explore the possibility of having a good model structure which better estimate the software effort. the tuning parameters for the GP evolutionary process is presented in Table VI. The Lisp expression developed using Lilgp1.1 program is given in Equation 7 and simplified in Equation 8. The convergence process for GP is presented in Figure (4). GP convergence to the best possible model with a good prediction capabilities. Table (VII) show the measured and estimated GP effort.

$$(-(+DL\ DL)\ (*(*(*(*0.022970.02588)M\ E)DL)M\ E)) \quad (7)$$

$$E = 2 \cdot DL - 0.59 \cdot 10^{-3} ME^2 \cdot DL \quad (8)$$

Parameter	Value
Max generations	100
Max tree depth	5
Max tree nodes	11
Initial tree depth	2-4
Crossover rate	0.8
Reproduction rate	0.1
Mutation rate	0.1
selection method	fitness_overselect

Table 4. LILGP Experimental Setupe For DL Based Model

5. Comparison with Other Models

The computed performance of the developed models is presented in Table VIII. The computed VAF is high in the case of the DL-ME based model and better than the case of the DL based model. This is an evidence that the inclusion in the ME as a variable in the modeling process of the effort enhance the model capabilities to better estimate the effort. Thus, GP was able to better find the function *f* which related the E and both DL and *ME*, *E = f(DL,ME)*.

In [13], authors presented an extended work on the use of Soft Computing Techniques to build a suitable model structure to utilize improved estimations of software effort for NASA software projects. A comparison between COCOMOPSO, Fuzzy Logic (FL), Halstead, Walston-Felix, Bailey- Basili and Doty models were provided. In Table IX, we show the MMRE criteria computed overall data set. It is shown that the GP and the COCOMO based PSO models have almost similar properties. The FL model is the model found to provide the minimum MMRE since it consists of three linear models with various membership functions. This gives an advantage of the FL model over other effort estimation models.

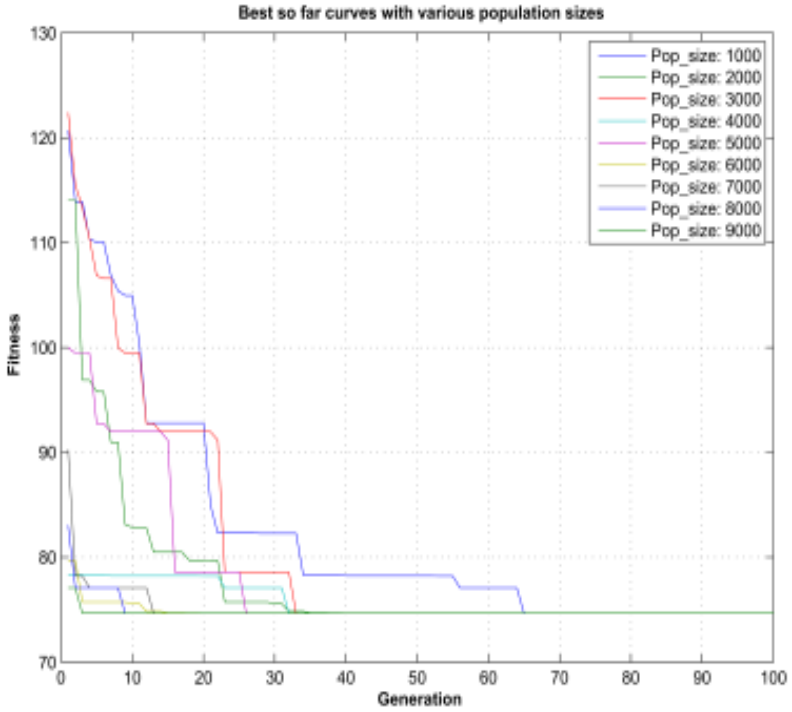

Figure 3. Convergence of GP with various population sizes for DL

Project No.	Measured Effort E	Estimated Effort $\hat{E}$
1	5.0000	3.6755
2	7.0000	5.4115
3	9.0000	7.3105
4	8.4000	8.6846
5	7.3000	13.4475
6	15.6000	16.6384
7	10.3000	17.9720
8	23.9000	21.2802
9	18.9000	21.7733
10	28.5000	35.7119
11	39.6000	50.2843
12	96.0000	71.4899
13	79.0000	71.8899
14	90.8000	82.2525
15	98.4000	97.8358
16	98.7000	109.9111
17	115.8000	121.3190
18	138.3000	130.6610

Table 5. Actual and Estimated Effort Using the GP Based DL Model

Parameter	Value
Max generations	100
Max tree depth	5
Max tree nodes	13
Initial tree depth	2-5
Crossover rate	0.8
Reproduction rate	0.1
Mutation rate	0.1
selection method	fitness overselect

Table 6.LILGP Experimental Setup For DL - ME Based Model

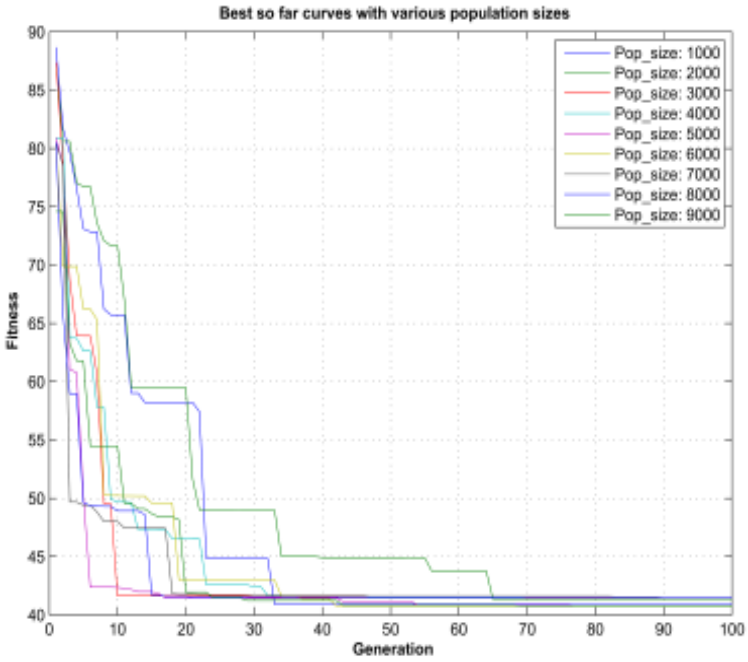

Figure 4. Convergence of GP with various population sizes for DL-ME

Project No.	Measured Effort E	Estimated Effort $\hat{E}$
1	5.0000	3.2213
2	7.0000	4.9542
3	9.0000	7.4987
4	8.4000	7.5003
5	7.3000	11.1440
6	15.6000	15.1964
7	10.3000	13.7844
8	23.9000	19.5830
9	18.9000	20.4562
10	28.5000	30.7175
11	39.6000	39.5524
12	96.0000	81.4143
13	79.0000	83.0210
14	90.8000	96.0407
15	98.4000	101.2538
16	98.7000	99.9621
17	115.8000	132.1414
18	138.3000	132.3302

Table 7. Actual and Esimated Effort Using GP Based DL-ME Model

Model	VAF	MMRE
DL based Model	96.5538	0.0052
DL-ME based Model	98.2346	0.0039

Table 8. The Computed Perormance of the Developed GP Models

Model	Fuzzy Model	GP Model	COCOMO based PSO	Walston-Felix Model	Bailey-Basili Model	Halstead Model	Doty Model
MMRE	0.0046	0.0052	0.0074	0.0822	0.0095	0.1479	0.1848

Table 9. The Computed MMRE Criterion for All Models Based DL Only

6. Conclusions and future work

In this paper we proposed a new model structure to estimate the software effort for projects sponsored by NASA using genetic programming. The performance of the developed GP model was tested on NASA software projects data presented in [34]. The developed software effort estimation model based GP was capable of providing good effort estimation as compared to other known model in the literature such as Halstead, Walston-Felix, Bailey-Basili and Doty models. the consideration of other attributes such as the Methodology while developing the effort most significantly improves the model prediction capabilities. GP was able to provide an advanced mathematical function utilizing the DL and ME such that the computed effort is more accurate.

References

[1] Kemere, C. F. (1987). An empirical validation of software cost estimation models, *Communication ACM*, 30, 416–429.

[2] Boehm, B. (1995). Cost Models for Future Software Life Cycle Process: COCOMO2. Annals of Software Engineering.

[3] Boehm, B. et al (2000). Software Cost Estimation with COCOMO II. Prentice Hall PTR.

[4] Menzies, T., Port, D., Chen, Z. Hihn, J., Stukes, S. (2005). Validation methods for calibrating software effort models, *In*: Proceedings of the 27th international conference on Software Engineering (ICSE'05), (New York, NY, USA), p. 587–595, ACM Press.

[5] Boehm, B (1981). Software Engineering Economics. Englewood Cliffs, NJ, Prentice-Hall.

[6] Lefley, M., Shepperd, M.J (2003). Using genetic programming to improve software effort estimation based on general data sets, *In*: GECCO'03: Proceedings of the 2003 International conference on Genetic and evolutionary computation, (Berlin, Heidelberg), p. 2477–2487, Springer-Verlag.

[7] Venkatachalam, A. R. (1993). Software cost estimation using artificial neural networks, *In*: Proceedings of 1993 IEEE International Conference on Neural Networks (ICNN'93*)*, V. 1, (Nagoya, Japan), p. 987–990, IEEE/INNS, Oct. 1993. University of New Hampshire.

[8] Sheta, A. F. (2006). Estimation of the COCOMO model parameters using genetic algorithms for NASA software projects, *Journal of Computer Science*, 2 (2) 118–123.

[9] Mittal, H., Bhatia, P (2007). A comparative study of conventional effort estimation and fuzzy effort estimation based on triangular fuzzy numbers, *International Journal of Computer Science and Security*, 1 (4) 36–47.

[10] Mittal, H., Bhatia, P (2007). Optimization criteria for effort estimation using fuzzy technique, *CLEI ELECTRONIC JOURNAL*, 10 (1) 1–11.

[11] Uysal, M. (2008). Estimation of the effort component of the software projects using simulated annealing algorithm, *In*: *World Academy of Science, Engineering and Technology*, 41, p. 258–261.

[12] Sandhu, P. S., Prashar, M., Bassi, P., Bisht, A. (2009). A model for estimation of efforts in development of software systems, *In*: *World Academy of Science, Engineering and Technology*, 56, p. 148–152.

[13] Sheta, A. Rine, D., Ayesh, A. (2008). Development of software effort and schedule estimation models using soft computing techniques, *In*: Proceedings of the 2008 IEEE Congress on Evolutionary Computation (IEEE CEC 2008) within the 2008 IEEE World Congress on Computational Intelligence (WCCI 2008), Hong Kong, 1-6 June, p. 1283–1289.

[14] Finnie, G. R., Wittig, G. W., Desharnais, J.-M. (1997). Estimating software development effort with case-based reasoning, *In*: Proceedings of the 2nd International Conference on Case-Based Reasoning (ICCBR-97) (D. B. Leake and E. Plaza, eds.), vol. 1266 of *LNAI*, (Berlin), p. 13–22, Springer, July 25–27.

[15] Shepperd, M. J, Schofield, C., Kitchenham, B.(1996). Effort estimation using analogy, *In*: ICSE, p. 170–178.

[16] Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., Webster, S. (2000). An investigation of machine learning based prediction systems, *The Journal of Systems and Software*, 53, p. 23–29, July.

[17]. Shukla, K. K (2000). Neuro-genetic prediction of software development effort, *Information & Software Technology*, 42, (10) 701–713.

[18] Langdon, W. B., Poli, R. (2002). Foundations of Genetic Programming. Springer-Verlag.

[19] Koza, J. R. (1992). Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press.

[20] Poli, R. Langdon, W. B. and McPhee, N. F. A field guide to genetic programming. Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, 2008. (With contributions by J. R. Koza).

[21] Hussian, A. Sheta, A. Kamel, M., Telbany, M., Abdelwahab, A. (2000). Modeling of a winding machine using genetic programming, *In:* Proceedings of the Congress on Evolutionary Computation (CEC2000), p. 398–402.

[22] Sheta, A., Gertler, J. (2000). Modeling the dynamics of an automotive engine using genetic programming, *In*: Proceedings of the International Symposium on Engineering of Natural and Artificial Intelligent Systems (ENAIS2001), American University in Dubai, U.A.E.

[23] Sheta, A., Mahmoud, A (2001). Forecasting using genetic programming, *In:* Proceedings of the 33 rd Southern Symposium on System Theory, March 19-20, Athens, Ohio, USA, p. 343–347.

[24] Al-Afeef, A. S. (2010). Image reconstructing in electrical capacitance tomography of manufacturing processes using genetic programming, Master's thesis, Al-Balqa Applied University, July.

[25] Al-Afeef, A., Alaa, F. S., Al-Rabea, A. (2010). Image reconstruction of a metal fill industrial process using genetic programming, *In: ISDA*, p. 12–17, IEEE.

[26] Davidson, J. W., Savic, D. A., Walters, G. A. (2001). Symbolic and numerical regression: experiments and applications, *In*: Developments in Soft Computing (R. John and R. Birkenhead, eds.), (De Montfort University, Leicester, UK), p. 175–182, Physica Verlag, 29-30 June.

[27] Xiong, S., Wang, W., Li, F. (2003). A new genetic programming approach in symbolic regression, *In:* Proceedings 15th IEEE International Conference on Tools with Artificial Intelligence, p. 161–165, IEEE, 3-5 Nov.

[28] Alaa, F. S., Al-Afeef, A (2010). A GP effort estimation model utilizing line of code and methodology for NASA software projects, *In*: ISDA, p. 290–295, IEEE.

[29] Kishore, J. K., Patnaik, L. M., Mani, V., Agrawal, V. K. (2000). Application of genetic programming for multicategory pattern classification, *IEEE Transactions on Evolutionary Computation*, 4, 242–258.

[30] Luke, S. (2000). Code growth is not caused by introns, *In*: Late Breaking Papers at the 2000 Genetic and Evolutionary Computation Conference (D. Whitley, ed.), (Las Vegas, Nevada, USA), p. 228–235, 8 July.

[31] Rouwhorst, S. E., Engelbrecht, A. P (2000). Searching the forest: Using decision trees as building blocks for evolutionary search in classification databases, *In*: Proceedings of the 2000 Congress on Evolutionary Computation CEC00, vol. 1, (La Jolla Marriott Hotel La Jolla, California, USA), p. 633–638, IEEE Press, 6-9 July.

[32] S. C. S. (Editor) (1992). Encyclopedia of Artificial Intelligence. John Wiley, 2 ed., January. 1792 p.

[33] Sheta, A (2006). Software effort estimation and stock market prediction using takagi-sugeno fuzzy models, *In*: *Proceedings of the 2006 IEEE Fuzzy Logic Conference, Sheraton, Vancouver Wall Centre, Vancouver, BC, Canada, July 16-21*, p.579–586.

[34] Bailey, J. W., Basili, V. R (1981). A meta model for software development resource expenditure, *In*: Proceedings of the International Conference on Software Engineering, p. 107–115, 1981.

[35] Zongker, D., Punch, B. (1996). "lilgp 1.01 user's manual," tech. rep., Michigan State University, USA, 26.