A Fuzzy Decision Tree to Estimate Development Effort for Web Applications

Ali Idri
Department of Software Engineering
ENSIAS, Mohammed Vth –Souissi University
BP. 713, Madinat Al Irfane, Rabat, Morocco

Sanaa Elyassami
Department of Software Engineering
ENSIAS, Mohammed Vth –Souissi University
BP. 713, Madinat Al Irfane, Rabat, Morocco

Abstract— **Web Effort Estimation is a process of predicting the efforts and cost in terms of money, schedule and staff for any software project system. Many estimation models have been proposed over the last three decades and it is believed that it is a must for the purpose of: Budgeting, risk analysis, project planning and control, and project improvement investment analysis. In this paper, we investigate the use of Fuzzy ID3 decision tree for software cost estimation, it is designed by integrating the principles of ID3 decision tree and the fuzzy set-theoretic concepts, enabling the model to handle uncertain and imprecise data when describing the software projects, which can improve greatly the accuracy of obtained estimates. MMRE and Pred are used, as measures of prediction accuracy, for this study. A series of experiments is reported using Tukutuku software projects dataset. The results are compared with those produced by three crisp versions of decision trees: ID3, C4.5 and CART.**

Keywords- Fuzzy Logic; Effort Estimation; Decision Tree; Fuzzy ID3; Software project.

I. INTRODUCTION

Estimation software project development effort remains a complex problem, and one which continues to attract considerable research attention. Improving the accuracy of the effort estimation models available to project managers would facilitate more effective control of time and budgets during software project development. Unfortunately, many software development estimates are quite inaccurate. Molokken and Jorgensen report in recent review of estimation studies that software projects expend on average 30-40% more effort than is estimated [13]. In order to make accurate estimates and avoid gross misestimations, several cost estimation techniques have been developed. These techniques may be grouped into two major categories: parametric models, which are derived from the statistical or numerical analysis of historical projects data [5], and non-parametric models, which are based on a set of artificial intelligence techniques such as artificial neural networks [9][4], case based reasoning [19], decision trees [20] and fuzzy logic [23][17]. In this paper, we are concerned with cost estimation models based on fuzzy decision trees especially Fuzzy Interactive Dichotomizer 3.

The decision tree method is widely used for inductive learning and has been demonstrating its superiority in terms of predictive accuracy in many fields [24][10]. The most widely used algorithms for building a decision tree are ID3 [11], C4.5 [12] and CART [14].

There are three major advantages when using estimation by decision trees (DT). First, decision trees approach may be considered as “white boxes”, it is simple to understand and easy to explain its process to the users, contrary to other learning methods. Second, it allows the learning from previous situations and outcomes. The learning criterion is very important for cost estimation models because software development technology is supposed to be continuously evolving. Third, it may be used to feature subset selection to avoid the problem of cost driver selection in software cost estimation model.

On the other hand, fuzzy logic has been used in software effort estimation. It's based on fuzzy set theory, which was introduced by Zadeh in 1965 [15]. Attempts have been made to rehabilitate some of the existing models in order to handle uncertainties and imprecision problems. Idri et al. [3] investigated the application of fuzzy logic to the cost drivers of intermediate COCOMO model while Pedrycz et al. [25] presented a fuzzy set approach to effort estimation of software projects.

In two earlier works [1][2] we have empirically evaluated the use of crisp decision tree techniques for software cost estimation. More especially, the two used crisp decision tree techniques are the ID3 and the C4.5 algorithms. The two studies are based on the COCOMO' 81 and a web hypermedia dataset. We have found that the decision tree designed with the ID3 algorithm performs better, in terms of cost estimates accuracy, than the decision tree designed with C4.5 algorithm for the two datasets.

The aim of this study is to evaluate and to discuss the use of fuzzy decision trees, especially the fuzzy ID3 algorithm in designing DT for software cost estimation.

Instead of crisp DT, fuzzy DT may allow to exploit complementary advantages of fuzzy logic theory which is the ability to deal with inexact and uncertain information when describing the software projects.

The remainder of this paper is organised as follows: In section II, we present the fuzzy ID3 decision tree for software cost estimation. The description of dataset used to perform the empirical studies and the evaluation criteria adopted to measure the predictive accuracy of the designed models are given in section III. Section IV focuses on the experimental design. In Section V, we present and discuss the obtained results when the fuzzy ID3 is used to estimate the software development effort. A comparison of the estimation results produced by means of the fuzzy ID3 model and three other crisp decision tree models is also provided in section V. A conclusion and an overview of future work conclude this paper.

II. Fuzzy ID3 For Software Cost Estimation

Based on the Concept Learning System algorithm, Quinlan proposed a decision tree called the Interactive Dichotomizer 3 (ID3). The ID3 technique is based on information theory and attempts to minimize the expected number of comparisons. The fuzzy ID3 is based on a fuzzy implementation of the ID3 algorithm [16][21]. It's formed of one root node, which is the tree top, or starting point, and a series of other nodes. Terminal nodes are leaves (effort). Each node corresponds to a split on the values of one input variable (cost drivers). This variable is chosen in order to reach a maximum of homogeneity amongst the examples that belong to the node, relatively to the output variable.

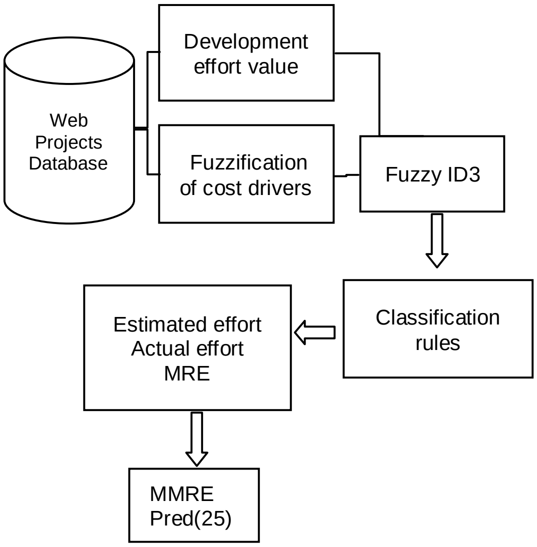


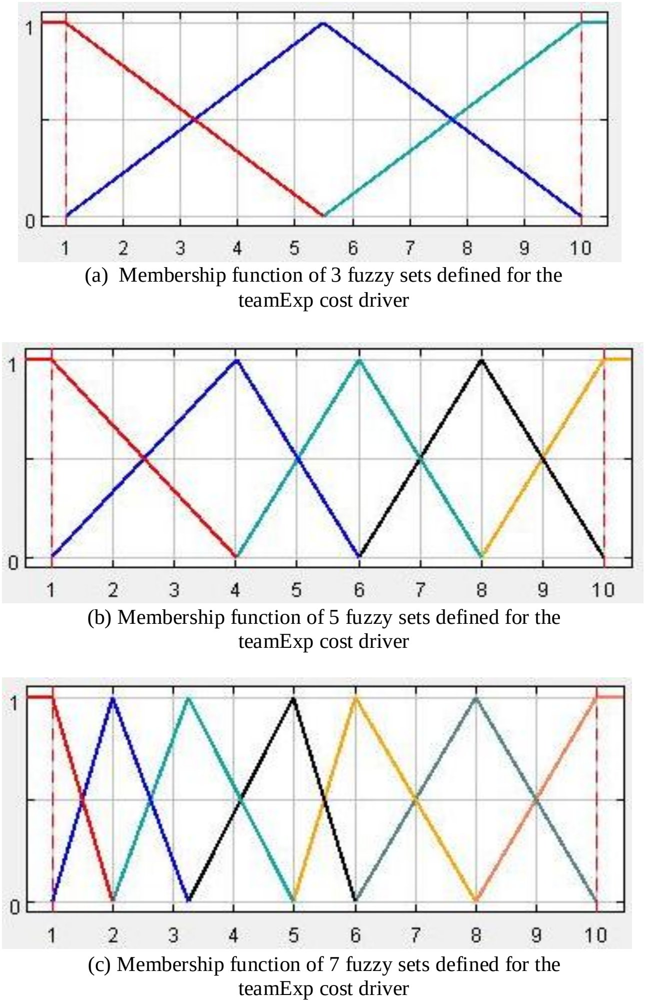
Figure 1. Fuzzy decision tree induction process

Fig. 1 illustrates the fuzzy decision tree induction process that consists on the fuzzification of the web cost drivers, the construction of the fuzzy decision tree, the prediction with the classification rules and the measure the accuracy of the estimates generated by the fuzzy ID3 decision tree.

The fuzzification of the software cost drivers converts crisp cost drivers into membership degrees to the different fuzzy sets of the partition. Many algorithms can be found in the specialized literature for generating partitions from data, we chose the Hierarchical Fuzzy Partitioning (HFP) [22]. It corresponds to an ascending procedure. At each step, for each given variable, two fuzzy sets are merged. This method combines two different clustering techniques, hierarchical clustering and fuzzy clustering techniques.

The triangular membership functions are used to represent the fuzzy sets because of its simplicity, easy comprehension, and computational efficiency.

Figure 2 illustrates the membership functions associated to the fuzzy sets of the team experience attribute.



(a) Membership function of 3 fuzzy sets defined for the teamExp cost driver

(b) Membership function of 5 fuzzy sets defined for the teamExp cost driver

(c) Membership function of 7 fuzzy sets defined for the teamExp cost driver

Figure 2. Membership functions associated to the fuzzy sets of the teamExp attribute

The fuzzy decision tree is interpreted by rules, Each path of the branches from root to leaf can be converted into a rule with condition part represents the attributes on the passing branches from root to the leaf and the conclusion part represents the class at the leaf of the form: IF (condition 1 and condition 2 .. and condition n) THEN C, where the conditions are extracted from the nodes and C is the leaf.

Fig. 3 illustrates an example of fuzzy ID3 decision tree for software development effort where MF represents the membership function used to define fuzzy sets for each cost driver.

III. Data Description and Evaluation Criteria

This section describes the dataset used to perform this empirical study and the evaluation criteria adopted to measure the estimates accuracy of the designed software cost estimation model based on fuzzy ID3 method.

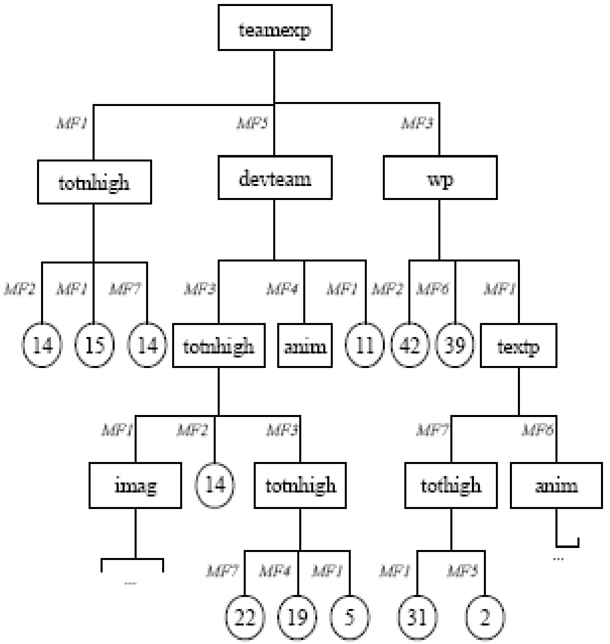


Figure 3. An example of fuzzy ID3 decision tree for software development effort

A. *Data Descriptions*

The Tukutuku dataset contains 53 web projects. [7] Each web application is described using 9 numerical attributes such as: the number of html or shtml files used, the number of media files and team experience (see Table I). However, each project volunteered to the Tukutuku database was initially characterized using more than 9 software attributes, but some of them were grouped together. For example, we grouped together the following three attributes: the number of new Web pages developed by the team, the number of Web pages provided by the customer and the number of Web pages developed by a third party (outsourced) in one attribute reflecting the total number of Web pages in the application (TotWP).

TABLE I. SOFTWARE ATTRIBUTES FOR THE TUKUTUKU DATASET

Attributes	*Description*
TeamExp	Average team experience with the development language(s) employed
DevTeam	Size of development team
TotWP	Total number of web pages
TextPages	Number text pages typed (~600 words)
TotImg	Total number of images
Anim	Number of animations
AV	Number of audio/video files
TotHigh	Total Number of high effort features/functions
TotNHigh	Total Number of low effort features/functions

B. *Evaluation criteria*

We employ the following criteria to measure the accuracy of the estimates generated by the fuzzy ID3. A common criterion for the evaluation of effort estimation models is the magnitude of relative error (MRE), which is defined as

$$MRE = \left| \frac{Effort_{actual} - Effort_{estimated}}{Effort_{actual}} \right| \qquad (1)$$

where $Effort_{actual}$ is the actual effort of a project in the dataset, and $Effort_{estimated}$ is the estimated effort that was obtained using a model or a technique.

The *MRE* values are calculated for each project in the datasets, while mean magnitude of relative error (MMRE) computes the average over *N* projects.

$$MMRE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{Effort_{actual,i} - Effort_{estimated,i}}{Effort_{actual,i}} \right| \times 100 \qquad (2)$$

The acceptable target values for *MMRE* are $MMRE \leq 25$. This indicates that on the average, the accuracy of the established estimation model would be less than 25%.

Another widely used criterion is the prediction *Pred(p)* witch represents the percentage of *MRE* that is less than or equal to the value p among all projects. This measure is often used in the literature and is the proportion of the projects for a given level accuracy [18]. The definition of *Pred(p)* is given as follows:

$$Pred(p) = \frac{k}{N} \qquad (3)$$

Where *N* is the total number of observations and *k* is the number of observations whose *MRE* is less or equal to *p*. A common value for *p* is 25, witch also used in the present study. The prediction at 25%, *Pred(25)*, represents the percentage of projects whose *MRE* is less or equal to 25%. The acceptable values for *Pred(25)* are $\Pr ed(25) \geq 75$.

IV. EXPERIMENT DESIGN

This section describes the experiment design of the fuzzy ID3 decision tree on the Tukutuku dataset. The Hierarchical Fuzzy Partitioning method is chosen for generating the partitions.

The use of fuzzy ID3 to estimate software development effort requires the determination of the parameters, namely the number of input variables, the maximum number of fuzzy sets for each input variable and the significant level value. The last two parameters play an essential role in the generation of fuzzy decision trees. It greatly affects the calculation of fuzzy entropy and classification results of Fuzzy Decision trees.

The number of input variables is the number of the attributes describing the historical software projects in the used dataset. Therefore, when applying fuzzy ID3 to Tukutuku dataset, the number of input variables is equal to 9. Concerning the significant level parameter, is the membership degree for an example to be considered as belonging to the node, is fixed to 0.2 for all experiments.

In the present paper we are interested in studying the impact of the number of fuzzy sets on the accuracy of fuzzy ID3. A series of experiments is conducted with the fuzzy ID3 algorithm each time using a different value of the fuzzy sets. The number of fuzzy sets is varied within the interval [3, 9].

V. Overview Of The Experimental Results

This section presents and discusses the results obtained when applying the fuzzy ID3 to the Tukutuku dataset. The calculations were made using Fispro software [8]. We conducted several experiments using different configurations of fuzzy ID3 obtained by varying the number of fuzzy sets. The aim is to determine which configuration improves the estimates.

The results for the different configurations have been compared. Fig. 4 and Fig. 5 show the accuracy of the fuzzy ID3 model, measured in terms of MMRE and Pred, on Tukutuku dataset.

Fig. 4 compares the accuracy of the model, in terms of MMRE, when varying the number of the fuzzy sets. We note that the fuzzy ID3 model generates a lower MMRE when increasing the number of fuzzy sets. For example, when setting the number of fuzzy sets at 9 the model produces a prediction error equal to 2.38 (MMRE=2.38) and when setting the number of fuzzy sets at 4 the model produces a prediction error equal to 52.19 (MMRE = 52.19).

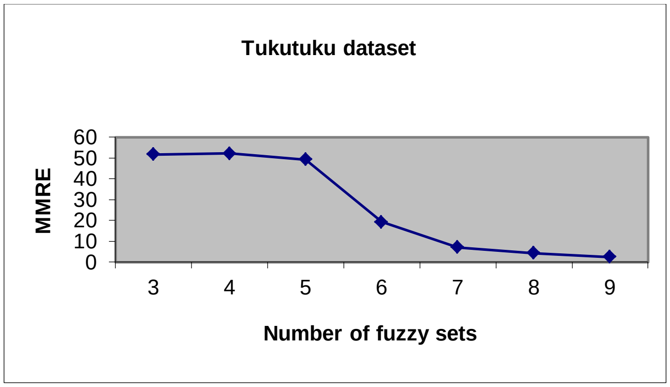


Fig. 4 Relationship between the accuracy of Fuzzy ID3 (MMRE) and the number of fuzzy sets

Fig. 5 shows the results of the model, in terms of Pred(25), when varying the number of the fuzzy sets. From this figure, we note that the accuracy of fuzzy ID3 model performs much better when increasing the number of the fuzzy sets and it's acceptable for the number of fuzzy sets greater than or equal to 6.

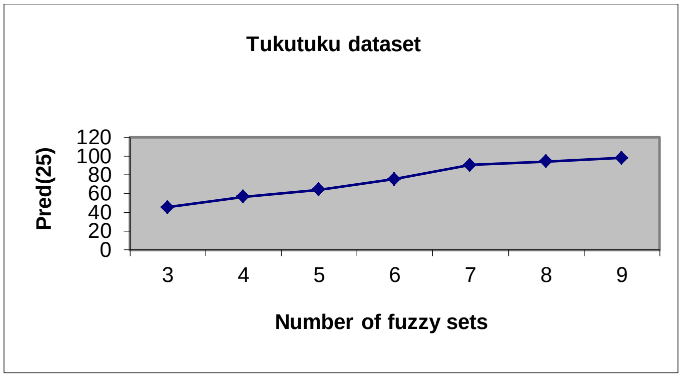


Figure 5. Relationship between the accuracy of Fuzzy ID3 (Pred) and the number of fuzzy sets

Table II summarizes the results obtained using different configurations of fuzzy ID3 for Tukutuku dataset. It shows the variation of the accuracy according to the number of fuzzy sets for Tukutuku dataset.

Table II. MMRE and Pred results of different fuzzy ID3 configurations for Tukutuku dataset

Number of fuzzy sets	MMRE	Pred(25)
3	51,68	45,28
4	52,19	56,6
5	49,3	64,15
6	19,27	75,47
7	7,09	90,57
8	4,3	94,34
9	2,38	98,11

The comparisons between the results produced by the fuzzy ID3 decision tree model and three other decision trees models: crisp ID3 decision tree model, C4.5 decision tree model [2] and CART model [6].

The best results obtained by means of the 4 models are compared in terms of MMRE and Pred(25). The comparison result is given in table III.

Table III. Result of the different models used on Tukutuku dataset

	Performance Criteria	
Decision tree models	Evaluation MMRE	Pred(25)
Crisp ID3	32	70
C4.5	28	70
CART	25	78
Fuzzy ID3	2,38	98,11

The experimental results show that the fuzzy ID3 model shows better estimation accuracy than the other crisp models in terms of MMRE and Pred(25).

For example, the improvement is 92.56% based on the fuzzy ID3 model MMRE and the crisp ID3 MMRE and is the 90.48% based on the fuzzy ID3 model MMRE and the CART model MMRE.

VI. CONCLUSION

In this paper, we have empirically studied a fuzzy ID3 model for software effort estimation. This fuzzy ID3 model is trained and tested using the tukutuku software projects dataset. The results show that the use of an optimal number of fuzzy sets improves greatly the estimates generated by fuzzy ID3 model. The comparison with the crisp decision tree models shows encouraging results.

To generalize this affirmation, we are looking currently in applying the fuzzy ID3 decision tree model on other historical software projects datasets.

REFERENCES

[1] A. Idri, S. Elyassami, Software Cost Estimation Using Decision Trees, In Proceeding of Sixième Conférence sur les Systèmes Intelligents: Théories et Applications (SITA'10), Rabat, Morocco, 4-5 Mai, 2010. pp. 120-125

[2] A. Idri, S. Elyassami, Web Effort Estimation Using Decision Trees, In Proceeding International Symposium on INnovations in Intelligent SysTems and Applications (INISTA 2010), Kayseri, Turkey, 21-24 Juin, 2010. pp. 224-228.

[3] A. Idri, L. Kjiri, and A. Abran, “COCOMO Cost Model Using Fuzzy Logic”, 7th International Conference on Fuzzy Theory & Technology, Atlantic City, NJ, February, 2000. pp. 219-223.

[4] A. Idri, and A. Abran, and S. Mbarki, "An Experiment on the Design of Radial Basis Function Neural Networks for Software Cost Estimation", in 2nd IEEE International Conference on Information and Communication Technologies: from Theory to Applications, 2006, Vol. 1, pp. 230-235.

[5] B.W. Boehm, Software Engineering Economics, Place: Prentice-Hall, 1981.

[6] E. Mendes, “Cost Estimation techniques for web projects”, 2008, pp. 203–239..

[7] B.A Kitchenham and E. Mendes, “A Comparison of Cross-company and Within-company Effort Estimation Models for Web Applications”, Proceedings of EASE Conference, 2004, pp. 47-56.

[8] Guillaume, S., Charnomordic, B., Lablee, J.-L., 2002. FisPro: Logiciel open source pour les systemes d'inference floue. http://www.inra.fr/bia/M/fispro. INRA-Cemagref.

[9] G. R. Finnie, and G. Witting, and J.-M. Desharnais, "A Comparison of Software Effort Estimation Techniques: Using Function Points with Neural Networks, Case-Based Reasoning and Regression Models", Systems and Software, Vol. 39, No. 3, 1997, pp. 281-289.

[10] H. Berger, D. Merkl, and M. Dittenbach, "Exploiting Partial Decision Trees for Feature Subset Selection in e-Mail Categorization," in Proceedings of the 2006 ACM Symposium on Applied Computing (SAC 2006), Dijon, France, 2006, pp. 1105-1109.

[11] J. R. Quinlan, “Induction on decision tree,” Machine Learning, Vol. 1, 1986, pp. 81-106.

[12] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[13] K. Molokken, and M. Jorgensen, "A Review of Surveys on Software Effort Estimation", in International Symposium on Empirical Software Engineering, 2003, pp. 223-231.

[14] L. Breiman, J.H. Friedman, R.A. Olsen & C.J. Stone. Classification and Regression Trees. Wadsworth, 1984.

[15] L. A. Zadeh, “Fuzzy sets”, Information and Control, vol 8, 1965, pp. 338–353.

[16] M. Umano, H. Okamoto, I. Hatono, H. Tamura, F. Kawachi, S. Umedzu, J. Kinoshita, “Fuzzy Decision Trees by Fuzzy ID3 algorithm and Its Application to Diagnosis Systems”, In Proceedings of the third IEEE Conference on Fuzzy Systems, vol. 3, Orlando, 1994, pp. 2113-2118.

[17] M. W. Nisar, and Y.-J. Wang, and M. Elahi, "Software Development Effort Estimation Using Fuzzy Logic – A Survey", in 5th International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 421-427.

[18] M. Korte and D. Port, “Confidence in Software Cost Estimation Results Based on MMRE and PRED”, PROMISE’08, May 12-13, 2008, pp . 63-70.

[19] M. Shepperd and C. Schofield. “Estimating Software Project Effort Using Analogies.” Transactions on Software Engineering, vol. 23, no. 12, 1997, pp. 736-747.

[20] R. W. Selby, and A.A. Porter, "Learning from examples: generation and evaluation of decision trees for software resource analysis", IEEE Transactions on Software Engineering, Vol. 14, No. 12, 1988, pp. 1743-1757.

[21] R. Weber, Fuzzy ID3: a class ofmethods for automatic knowledge acquisition, Proceedings ofthe 2nd International Conference on Fuzzy Logic and Neural Networks, Iizuka, Japan, July 17–22, 1992, pp. 265–268.

[22] S. Guillaume and B. Charnomordic, “Generating an interpretable family of fuzzy partitions”, IEEE Transactions on Fuzzy Systems, 12 (3), June 2004, pp. 324– 335.

[23] V. Sharma, and H. K. Verma, "Optimized Fuzzy Logic Based Framework for Effort Estimation in Software Development", Computer Science Issues, Vol. 7, Issue 2, No. 2, 2010, pp. 30-38.

[24] W. Pedrycz and Z. A. Sosnowski, “The design of decision trees in the framework of granular data and their application to software quality models,” Fuzzy Sets and Systems, Vol. 1234, 2001, pp. 271-290.

[25] W. Pedrycz, J.F. Peters, S. Ramanna, A Fuzzy Set Approach to Cost Estimation of Software Projects, Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering Shaw Conference Center , Edmonton Alberta, Canada. 1999, pp. 1068-1073.

AUTHORS PROFILE

A. Idri is a Professor at Computer Science and Systems Analysis School (ENSIAS, Rabat, Morocco). He received DEA (Master) (1994) and Doctorate of 3rd Cycle (1997) degrees in Computer Science, both from the University Mohamed V of Rabat. He has received his Ph.D. (2003) in Cognitive Computer Sciences from ETS, University of Quebec at Montreal. His research interests include software cost estimation, software metrics, fuzzy logic, neural networks, genetic algorithms and information sciences.

S. Elyassami received her engineering degree in Computer Science from the UTBM, Belfort-Montbeliard, France, in 2006. Currently, she is preparing her Ph.D. in computer science in ENSIAS. Her research interests include software cost estimation, soft