S. D. Shtovba, Dc. Sc. (Eng.), Prof.; A. A. Yakovenko

PREDICTION OF SOFTWARE SYSTEM DEVELOPMENT EFFORTS USING A FUZZY HYBRID MODEL

For prediction of software system development efforts the paper proposes application of a fuzzy hybrid model in which rule antecedents are given by fuzzy terms and consequents – by "inputs – output" linear dependencies with fuzzy coefficients. According to experimental data, identification of the dependence of software module development efforts on the programmer's experience, novelty and complexity of the task has been performed. It has been found that the proposed hybrid model provides an improved accuracy as compared with other five competitive models.

Keywords: *fuzzy hybrid model, fuzzy inference, fuzzy regression, fuzzy identification, software system, effort prediction.*

Introduction

In the creation of software systems there arises a problem of their development effort estimation. As a rule, the entire development process is divided into stages, each of them consisting of definite tasks. Efforts put into each task are estimated by the leader of the developers' team. When the efforts are estimated, it is rather difficult to take into account all the influencing factors, which results in a significant prediction error and degrades the project management quality.

There are many models of predicting software system development efforts [1, 2]. The most popular among them is a 22-factor model COCOMO II – Constructive Cost Model [3]. The results of practical application of these models show that they do not fully take into account all the features of software system development process. The difficulties of estimation are caused by the uncertainty of source data and the parameters of the effort estimation models due to significant influence of the human factor. Therefore, there is an interest in the application of new prediction methods that are well adapted for taking into account such uncertainty, e. g. fuzzy identification technologies. The paper aims at verification of the possibility to predict software system development efforts using a new fuzzy hybrid model [4]. This model consists of inference rules, the antecedents of which are given by fuzzy sets and consequents – by fuzzy regression equations. Such hybrid format makes it possible to describe a complex dependence only by several rules that take into account the uncertainty of the rule action boundaries by means of fuzzy antecedents. Simultaneously, the uncertain power of influence of the factors is taken into account by means of fuzzy coefficients in regression dependencies that make up the consequents of the rules. Uncertainty of the initial data is taken into account by presenting them in the form of fuzzy sets with further inference for fuzzy values of the influencing factors.

1. Fuzzy hybrid rule base

Let us write the hybrid rule base in the following way [4]:

$$\text{If } \left(x_1=\tilde{a}_{j1} \text{ and } x_2=\tilde{a}_{j2} \text{ and ... and } x_n=\tilde{a}_{jn}\right)\text{, then } y=\tilde{d}_j\,,\ j=\overline{1,m}\,, \tag{1}$$

where $\tilde{a}_{ji}$ – fuzzy term that evaluates input variable x_i in the *j*-th rule $i=\overline{1,n}$, $j=\overline{1,m}$;

m – the number of rules;

$\tilde{d}_j=\tilde{k}_{j0}+\tilde{k}_{j1}x_1+\tilde{k}_{j2}x_2+...+\tilde{k}_{jn}x_n$ – consequent of the *j*-th rule represented by a linear function with fuzzy coefficients $\tilde{k}_{j0},\ \tilde{k}_{j1},...,\tilde{k}_{jn}$.

Unlike Sugeno knowledge base [5], in (1) coefficients in the consequents of the rules are given by fuzzy numbers. Therefore, an expert can describe these fuzzy coefficients by such linguistic estimates as “little influence”, “moderate influence”, “strong influence”, etc., which reflect his knowledge about the degree of influence of the corresponding input variable on the output variable. In accordance with such linguistic estimates, a core of the fuzzy coefficient in (1) can be determined. Spreading of the fuzzy coefficient depends on the expert confidence in his knowledge, which can be expressed by the following terms: “absolutely reliable”, “almost reliable”, “more or less reliable”, etc. The more reliable the knowledge, the more concentrated the membership function of the fuzzy coefficient is.

2. Inference on the hybrid fuzzy rule base

Let us perform inference on rule base (1) in the following way. First, for the vector $X^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right)$ of current values of the input variables we calculate fuzzy values of the consequents according to the rules of fuzzy arithmetic:

$$\widetilde{d}_j = \widetilde{k}_{j0} + \widetilde{k}_{j1}x_1^* + \widetilde{k}_{j2}x_2^* + \ldots + \widetilde{k}_{jn}x_n^*, \; j = \overline{1, m}. \tag{2}$$

This will transform (1) into Mamdani knowledge base. Therefore, further steps will be performed according to Mamdani algorithm [6]. It should be noted, that for each input vector Mamdani knowledge base with a unique set of fuzzy consequents is created.

In accordance with Mamdani algorithm, we calculate the degree of fulfillment of the j-th rule antecedent for the input vector $X^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right)$ as follows:

$$\mu_j(X^*) = \mu_{j1}(x_1^*) \;\wedge\; \mu_{j2}(x_2^*) \;\wedge \ldots \wedge\; \mu_{jn}(x_n^*)\;, \; j = \overline{1, m}, \tag{3}$$

where $\mu_{ji}(x_i^*)$ – the degree of membership of the input value x_i^* to fuzzy term $\widetilde{a}_{ij}$, $i = \overline{1, n}$;

$\wedge$ – a norm that in Mamdani algorithm is typically implemented by the operation of minimum or by a product.

As a result of inference on the j-th rule of the base, we obtain the following fuzzy set:

$$\widetilde{d}_j^* = imp\left(\widetilde{d}_j, \; \mu_j(X^*)\right), \quad j = \overline{1, m}, \tag{4}$$

where *imp* denotes implication that is realized by the operation of minimum.

Geometric interpretation of the implication consists in cutting the graph of membership function $\mu_{d_j}(y)$ of fuzzy consequent (2) at the level $\mu_j(X^*)$:

$$\widetilde{d}_j^* = \int\limits_{y \in [\underline{y}, \overline{y}]} \min\left(\mu_j(X^*), \; \mu_{d_j}(y)\right) / y,$$

where $[\underline{y}, \overline{y}]$ is variation range of the output variable y.

The result of inference on all the rules is found by aggregation of fuzzy sets (4):

$$\widetilde{y}^* = agg\left(\widetilde{d}_1^*, \; \widetilde{d}_2^*, \ldots, \widetilde{d}_m^*\right), \tag{5}$$

where *agg* – aggregation of the fuzzy sets that is realized by the operation of maximum applied to the membership functions.

Crisp resulting value y^* is determined through defuzzification of fuzzy set $\widetilde{y}^*$ by the centroid method.

If source data are given by fuzzy values $\widetilde{X}^* = \left(\widetilde{x}_1^*, \widetilde{x}_2^*, \ldots, \widetilde{x}_n^*\right)$, then we calculate fuzzy values of the consequents using the rules of fuzzy arithmetic: $\widetilde{d}_j = \widetilde{k}_{j0} + \widetilde{k}_{j1}\widetilde{x}_1^* + \widetilde{k}_{j2}\widetilde{x}_2^* + \ldots + \widetilde{k}_{jn}\widetilde{x}_n^*, \; j = \overline{1, m}$.

We will perform further inference by formulas (3) – (4), only instead of $\mu_{ji}(x_i^*)$ we use $\mu_{ji}(\tilde{x}_i^*)$ – the degree of membership of fuzzy input value $\tilde{x}_i^*$ to fuzzy term $\tilde{a}_{ij}$, $j=\overline{1,m}$, $i=\overline{1,n}$. We calculate these values in the following way [6]:

$$\mu_j(\tilde{x}_i^*) = height\left(\tilde{x}_i^* \cap \tilde{a}_{ij}\right), \tag{5}$$

where $height$ is the height of fuzzy set.

3. Parametric identification of the dependencies using a fuzzy hybrid rule base

Let us write a training set from M "inputs - output" pairs as follows:

$$\left(X_r, y_r\right), \qquad r=\overline{1,M},$$

where X_r – input vector in the *r*-th line of the sample and y_r – the corresponding output.

Let us denote the model on the basis of fuzzy hybrid rule base (1) with parameters P as $y = F(P, X)$. Parametric identification involves finding vector *P* that provides:

$$RMSE = \sqrt{\frac{1}{M}\sum_{r=\overline{1,M}}\left(y_r - F\left(P, X_r\right)\right)^2} \to \min. \tag{6}$$

In (7) controlled variables *P* correspond to the parameters of membership functions of fuzzy sets from antecedents and consequents of rules (1). To preserve interpretability of the model, constraints are imposed on the parameters of fuzzy sets $\tilde{a}_{ij}$, $i=\overline{1,n}$, $j=\overline{1,m}$ in accordance with [7].

4. Experimental data for identification of the effort prediction model

To synthesize a model for predicting software system development efforts, we use the data provided by "Orneon" company. They refer to the project on creating the game of the "Quest" type. The output variable is y – effort put into the task, measured in man / days. The input variables include the following factors:

x_1 – programmer's experience in months;

x_2 – the novelty of the task that is evaluated on a scale from 1 to 10;

x_3 – the complexity of the task that is evaluated on a scale from 1 to 10.

Training set is formed from the tasks that were performed during the first three months of the project realization and the test sample – from later tasks. The training set includes 107 cases (Fig. 1) and the test set – 106 (Fig. 2).

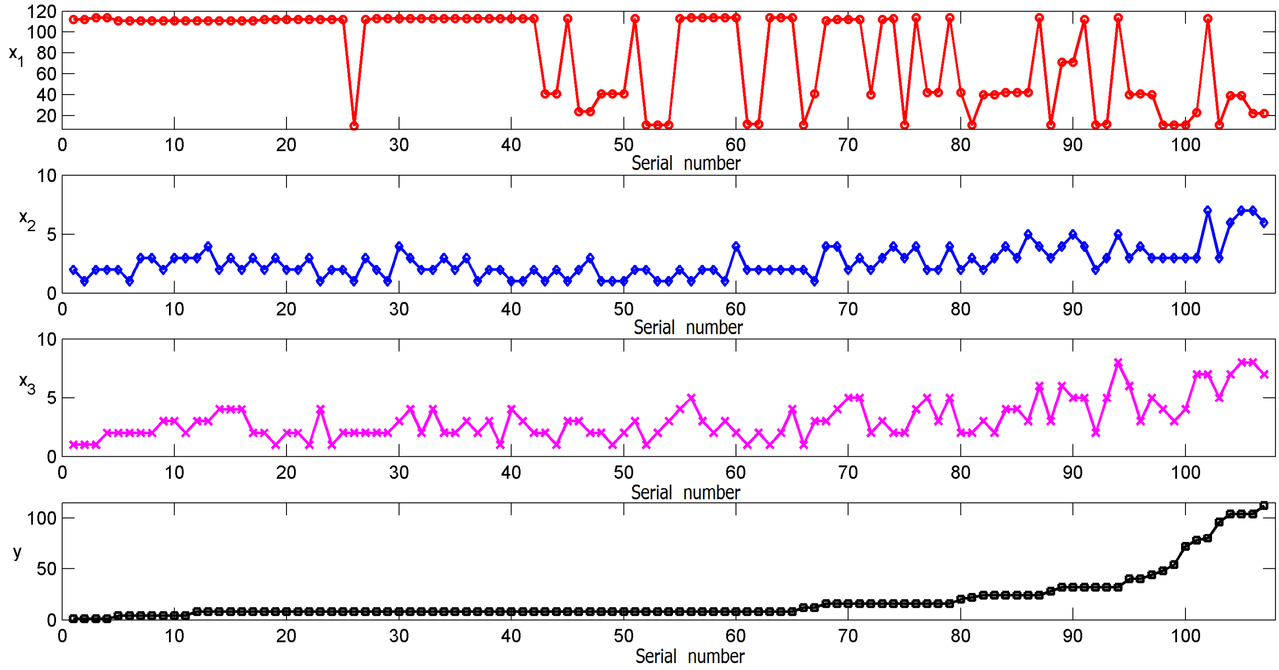

Fig. 1. The training set

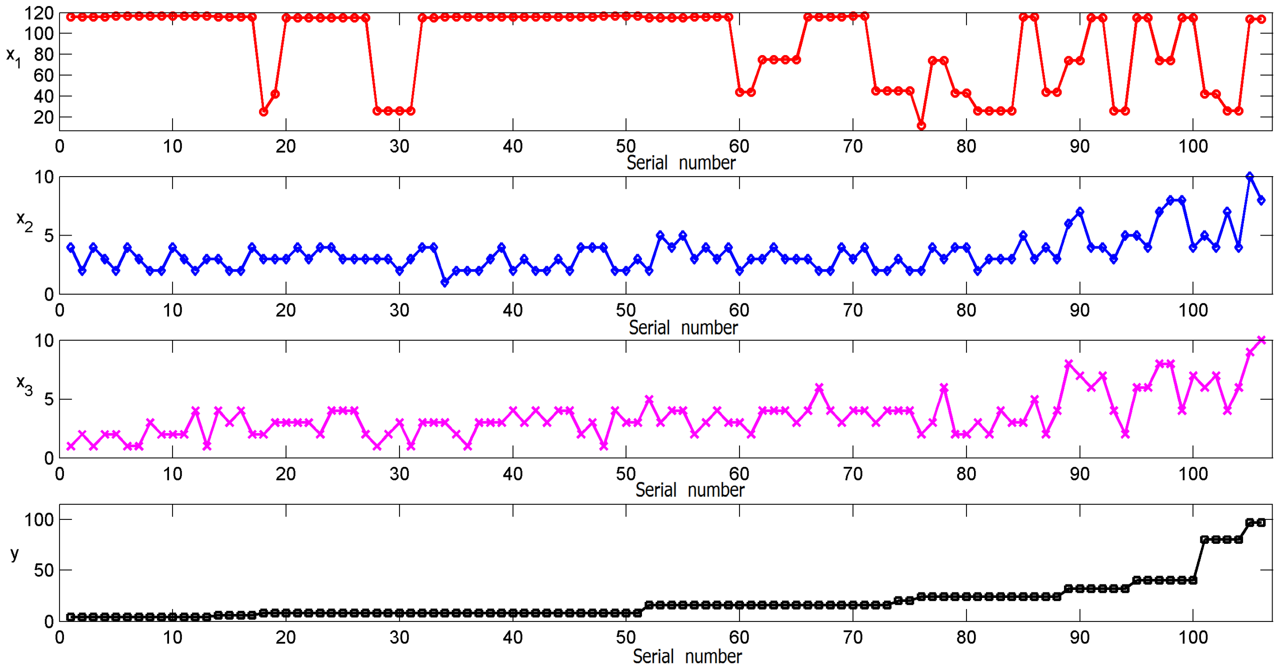

Fig. 2. The test set

5. Models of predicting software system development efforts

We perform modeling of the dependence $y = f(x_1, x_2, x_3)$ using the following three hybrid rules:

If $x_1 = Low$, then $y = \tilde{k}_{10} + \tilde{k}_{11}x_1 + \tilde{k}_{12}x_2 + \tilde{k}_{12}x_3$;

If $x_1 = Average$, then $y = \tilde{k}_{20} + \tilde{k}_{21}x_1 + \tilde{k}_{22}x_2 + \tilde{k}_{22}x_3$;

If $x_1 = High$, then $y = \tilde{k}_{30} + \tilde{k}_{31}x_1 + \tilde{k}_{32}x_2 + \tilde{k}_{32}x_3$.

In these rules the terms *Low* and *Average* as well as fuzzy coefficients of the consequents are given by triangular membership functions. A triangular membership function includes three parameters (a, b, c) that correspond to the support (a, c) and the core (b) of the fuzzy set. Term $High$ is described by a trapezoidal membership function with four parameters (a, b, c, d) that correspond to the support (a, d) and the core of the fuzzy set. Parameters of the membership functions of the terms and fuzzy coefficients after training are summarized in Table 1.

Table 1. Parameters of the membership functions of the terms and fuzzy coefficients of the fuzzy hybrid model

Fuzzy set	Parameters of the membership functions
$\tilde{k}_{10}$	(-3.13, 0.87, 4.97)
$\tilde{k}_{11}$	(-4.84, -3.26, -2.86)
$\tilde{k}_{12}$	(7.29, 9.22, 10.71)
$\tilde{k}_{13}$	(11.54, 12.74, 14.04)
$\tilde{k}_{20}$	(-5.95, -4.91, -3.97)
$\tilde{k}_{21}$	(-1.41, -1.37, -1.37)
$\tilde{k}_{22}$	(3.97, 8.32, 10.12)
$\tilde{k}_{23}$	(7.99, 14.97, 16.9)
$\tilde{k}_{30}$	(-2.14, 1.88, 8.09)
$\tilde{k}_{31}$	(-0.46, -0.46, -0.23)
$\tilde{k}_{32}$	(9.78, 10.7, 11.55)
$\tilde{k}_{33}$	(2.11, 2.9, 5.29)
Low	(0, 0, 18.59)
$Average$	(-10.91, 43.42, 78.17)
$High$	(22.6, 54.64, 120, 138)

In order to compare identification quality, we synthesized 5 competitive models:

- linear

$$y = -1 - 0.22x_1 + 6.84x_2 + 6.15x_3;$$

- quadratic

$$y = 24.51 - 0.51x_1 - 1.38x_2 + 0.46x_3 + 0.002x_1^2 + 1.165x_2^2 + 0.689x_3^2;$$

- polynomial of degree 1/2

$$y = 84.1 + 0.12x_1 + 18.56x_2 + 17.4x_3 - 4.82\sqrt{x_1} - 41.39\sqrt{x_2} - 42.1\sqrt{x_3};$$

- Wiener series

$$\begin{aligned} y \; = \; & 0.99 + 0.87x_1 + 0.95x_2 + 0.96x_3 - 0.026x_1^2 - 0.15x_1x_2 + 0.052x_1x_3 + 0.763x_2^2 - \\ - \; & 0.792x_2x_3 + 0.823x_3^2 - 0.0002x_1^3 - 0.0001x_1^2x_2 - 0.0004x_1^2x_3 + 0.0217x_1x_2^2 + \\ + \; & 0.0112x_1x_2x_3 - 0.018x_1x_3^2 - 0.0929x_2^3 - 0.0527x_2^2x_3 - 0.0527x_2x_3^2 + 0.0674x_3^3 \end{aligned};$$

- Sugeno fuzzy rule base

If $x_1 = Low$, then $y = 22.56 - 6.92x_1 + 10.33x_2 + 12.8x_3$;

If $x_1 = Average$, then $y = 2.05 - 0.79x_1 + 9.33x_2 + 12.11x_3$;

$$\text{If } x_1 = High, \text{ then } y = 0.97 - 0.1x_1 + 4.81x_2 + 3.03x_3 .$$

Parameters of the membership functions of the terms of Sugeno rule antecedents are summarized in Table 2.

Table 2. Parameters of the membership functions of the terms and fuzzy coefficients of the fuzzy hybrid model

Fuzzy set	Membership function parameters
Low	(0, 0, 29.13)
Average	(-11.49, 28.07, 67.63)
High	(21.3, 54.8, 129.68, 138)

Comparison of the modeling results with the experimental data are presented in Fig. 3 and 4. According to competitive models, for certain cases labor effort less than 1 is predicted. For these cases the output value is set to 1. From Fig. 3 and 4 it is evident that the fuzzy hybrid model is the best one both by root mean squared error ($RMSE$) and by maximal absolute error ($MaxErr$). The model based on Sugeno rules showed similar results as it could be considered to be a particular case of a fuzzy hybrid model in which all coefficients in the consequents are given by crisp numbers.

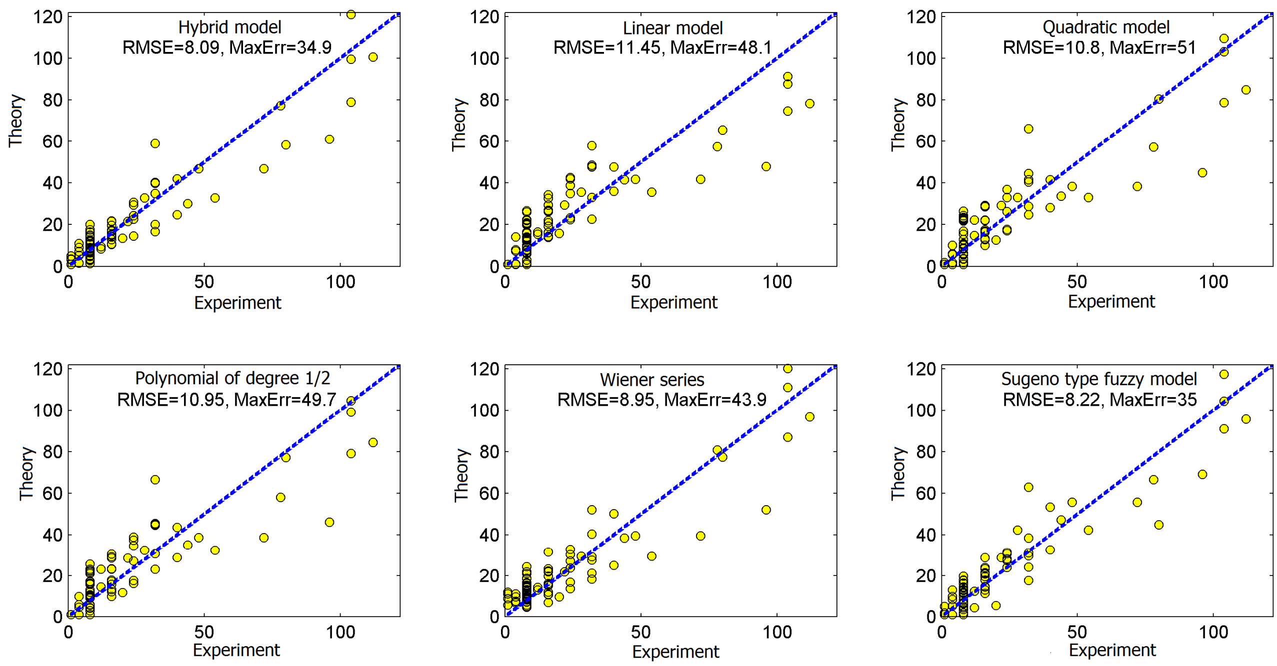

Fig. 3. Verification of the models on the training set

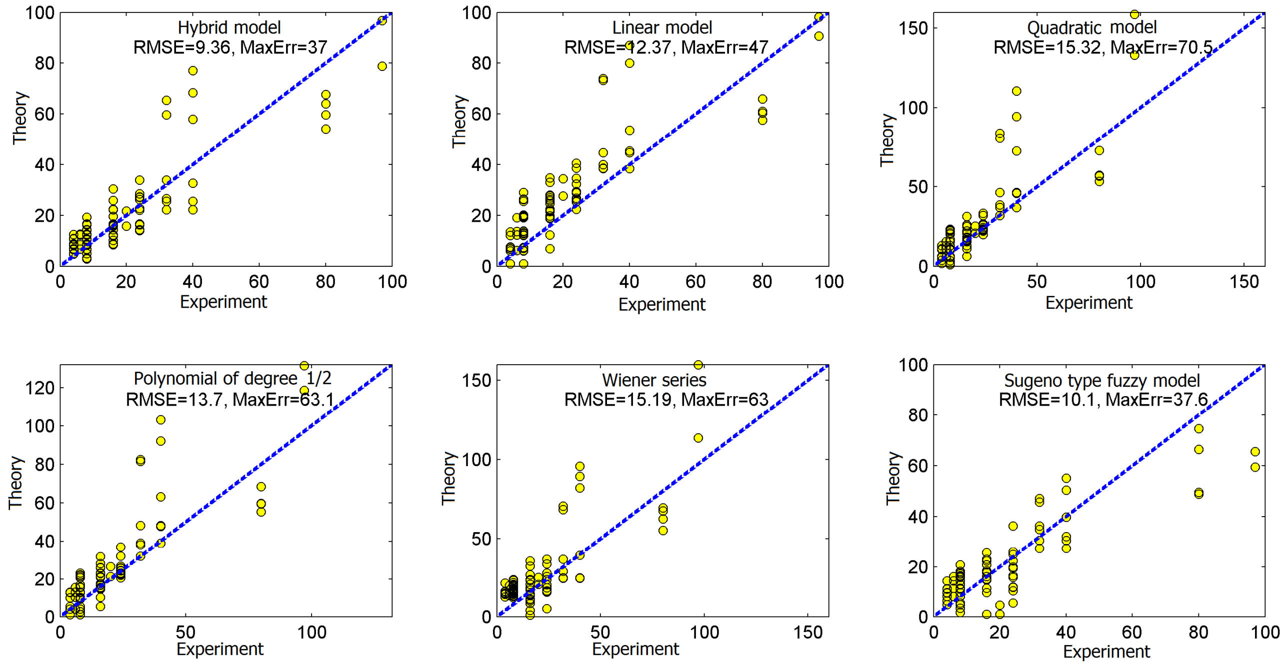

Fig. 4. Verification of the models on the test set

Conclusions

Application of a fuzzy hybrid model for predicting software system development efforts has been investigated. Comparison of the results with alternative models has been performed. The hybrid fuzzy model has shown the best results among the models considered. Therefore, its application for predicting time and effort spent on software system development is expedient as this will make it possible to improve project planning.

REFERENCES

1. Липаев В. В. Программная инженерия. Методологические основы : Учеб. / Липаев В. В. – Гос. ун-т – Высшая школа экономики. – М.: ТЕИС, 2006. – 608 с.

2. Hernandez-Lopez A. Software engineering job productivity – a systematic review / A. Hernandez-Lopez , R. Colomo-Palacios, A. Garcia-Crespo // International Journal of Software Engineering and Knowledge Engineering. – 2013. – Vol. 23. – №3. – P. 387 – 406.

3. Boehm B. W. Software Cost Estimation with COCOMO II. / Boehm B. W. et al. – New Jersey: Prentice Hall, 2000. – 502 p.

4. Штовба С. Д. Моделювання залежностей за допомогою нечіткої бази знань з нечіткими регресійними рівняннями / С. Д. Штовба / Вісник Вінницького політехнічного інституту. – 2011. – №3. – С. 195 – 199.

5. Takagi T. Fuzzy Identification of Systems and Its Applications to Modeling and Control / T. Takagi, M. Sugeno // IEEE Trans. on Systems, Man, and Cybernetics. – 1985. – Vol. 15. - №1. – P. 116 - 132.

6. Штовба С. Д. Проектирование нечетких систем средствами MATLAB / Штовба С. Д. – М.: Горячая линия – Телеком, 2007. – 288 с.

7. Штовба С. Д. Обеспечение точности и прозрачности нечеткой модели Мамдани при обучении по экспериментальным данным / С. Д. Штовба // Проблемы управления и информатики. – 2007. – №4. – С. 102 – 114.

Shtovba Serhiy – Professor, Dc. Sc. (Eng.), Professor of the Department of Computer Control Systems.

Yakovenko Anton – Student of the Institute of Automatics, Electronics and Computer Control Systems (InAECCS)

Vinnytsia National Technical University.