УДК 004.415

С. Д. Штовба, д. т. н., проф.; А. А. Яковенко

ПРОГНОЗИРОВАНИЕ ТРУДОЕМКОСТИ РАЗРАБОТКИ ПРОГРАММНЫХ СИСТЕМ С ПОМОЩЬЮ НЕЧЕТКОЙ ГИБРИДНОЙ МОДЕЛИ

Для прогнозирования трудоемкости разработки программных систем предложено использование нечеткой гибридной модели, в которой антецеденты правил задаются нечеткими термами, а консеквенты – линейными зависимостями "входы – выход" с нечеткими коэффициентами. По экспериментальным данным проведена идентификация зависимости трудоемкости разработки программных модулей от стажа программиста, новизны и сложности задачи. Установлено, что предложенная нечеткая гибридная модель обеспечивает более высокую точность по сравнению с другими пятью конкурентными моделями.

***Ключевые слова**: нечеткая гибридная модель, нечеткий вывод, нечеткая регрессия, нечеткая идентификация, программная система, прогнозирование трудоемкости.*

Введение

При создании программных систем возникает проблема оценки трудоемкости их разработки. Как правило, весь процесс разработки разбивают на этапы, каждый из которых состоит из конкретных задач. Трудоемкость выполнения каждой задачи оценивает лидер команды разработчиков. При оценке трудоемкости довольно трудно адекватно учесть все влияющие факторы, что приводит к существенной погрешности прогнозирования и ухудшает качество управления проектом.

Существует много моделей прогнозирования трудоемкости разработки программных систем [1, 2], среди которых одной из самых популярных является 22-факторная модель COCOMO II – Constructive Cost Model [3]. Результаты практического применения этих моделей указывают, что они не в полной мере учитывают все особенности процесса разработки программных систем. Трудности оценки обусловлены неопределенностью исходных данных и параметров моделей прогнозирования трудоемкости, что связано с существенным влиянием человеческого фактора, поэтому возникает заинтересованность в применении новых методов прогнозирования, которые хорошо приспособлены для учета такой неопределенности, например, технологий нечеткой идентификации.

Цель статьи заключается в проверке возможности прогнозирования трудоемкости разработки программных систем с помощью новой нечеткой гибридной модели [4]. Эта модель состоит из продукционных правил, антецеденты которых заданы нечеткими множествами, а консеквенты – нечеткими регрессионными уравнениями. Такой гибридный формат позволяет описать сложную зависимость лишь несколькими правилами, которые учитывают неопределенность границ действия правил с помощью нечетких антецедентов. Одновременно неопределенность силы влияния факторов учитывают с помощью нечетких коэффициентов в регрессионных зависимостях, составляющих консеквенты правил. Неопределенность исходных данных учитывают путем представления их в форме нечетких множеств с последующим логическим выводом для нечетких значений влияющих факторов.

1. База нечетких гибридных правил

Базу нечетких гибридных правил запишем таким образом [4]:

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

где $\tilde{a}_{ji}$ – нечеткий терм, которым оценена входная переменная x_i в j-м правиле $i=\overline{1,n}$, $j=\overline{1,m}$;

m – количество правил;

$\tilde{d}_j = \tilde{k}_{j0} + \tilde{k}_{j1}x_1 + \tilde{k}_{j2}x_2 + ... + \tilde{k}_{jn}x_n$ – консеквент j-го правила, представленный линейной функцией с нечеткими коэффициентами $\tilde{k}_{j0},\ \tilde{k}_{j1}, ..., \tilde{k}_{jn}$.

В отличие от базы знаний Сугено [5], в (1) коэффициенты в консеквентах правил заданы нечеткими числами. Поэтому эксперт может описать эти нечеткие коэффициенты лингвистическими оценками "мало влияет", "умеренно влияет", "сильно влияет" и т. д., отражающими его знания о степени влияния соответствующей входной переменной на выходную. По этим лингвистическими оценками можно определить ядро нечеткого коэффициента в (1). Размытость нечеткого коэффициента зависит от уверенности эксперта в достоверности своих знаний, которую можно выразить термами "абсолютно достоверно", "почти достоверно", "более-менее достоверно" и т. п. Чем достовернее знания, тем концентрированнее функция принадлежности нечеткого коэффициента.

2. Логический вывод по базе нечетких гибридных правил

Логический вывод по базе правил (1) осуществим следующим образом. Сначала для вектора $X^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right)$ текущих значений входных переменных по правилам нечеткой арифметики рассчитаем нечеткие значения консеквентов:

$$\tilde{d}_j = \tilde{k}_{j0} + \tilde{k}_{j1}x_1^* + \tilde{k}_{j2}x_2^* + ... + \tilde{k}_{jn}x_n^*,\ j=\overline{1,m}. \qquad (2)$$

Это преобразует (1) в базу правил Мамдани, поэтому дальнейшие шаги осуществим по алгоритму Мамдани [6]. Заметим, что для каждого входного вектора создают базу правил Мамдани с уникальным набором нечетких консеквентов.

По алгоритму Мамдани степень выполнения антецедента j-го правила для входного вектора $X^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right)$ рассчитаем так:

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

где $\mu_{ji}(x_i^*)$ – степень принадлежности входного значения x_i^* нечеткому терму $\tilde{a}_{ij}$, $i=\overline{1,n}$;

$\wedge$ – норма, которую в алгоритме Мамдани обычно реализуют операцией минимума или произведением.

В результате вывода по j-му правилу базы знаний получим такое нечеткое множество:

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

где *imp* обозначает импликацию, которую реализуют операцией минимума.

Геометрической интерпретацией импликации является "срезания" графика функции принадлежности $\mu_{d_j}(y)$ нечеткого консеквента (2) по уровню $\mu_j(X^*)$:

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

где $\left[\underline{y},\ \overline{y}\right]$ – диапазон изменения выходной переменной y.

Результат вывода по всем правилам находим агрегированием нечетких множеств (4):

$$\tilde{y}^* = agg\left(\tilde{d}_1^*,\ \tilde{d}_2^*, ..., \tilde{d}_m^*\right), \qquad (5)$$

где agg – агрегирование, которое реализуют операцией максимума над функциями принадлежности.

Четкое значение y^* определяем через дефаззификацию нечеткого множества $\tilde{y}*$ по методу центра тяжести.

Если нечеткие данные заданы нечеткими значениями $\tilde{X}^* = \left(\tilde{x}_1^*, \tilde{x}_2^*, \ldots, \tilde{x}_n^*\right)$, то по правилам нечеткой арифметики рассчитаем нечеткие значения консеквентов $\tilde{d}_j = \tilde{k}_{j0} + \tilde{k}_{j1}\tilde{x}_1^* + \tilde{k}_{j2}\tilde{x}_2^* + \ldots + \tilde{k}_{jn}\tilde{x}_n^*$, $j = \overline{1, m}$. Дальнейший вывод осуществим по формулам (3) – (4), только вместо $\mu_{ji}(x_i^*)$ используем $\mu_{ji}(\tilde{x}_i^*)$ – степень принадлежности нечеткого входного значения $\tilde{x}_i^*$ нечеткому терму $\tilde{a}_{ij}$, $j = \overline{1, m}$, $i = \overline{1, n}$. Эти значения рассчитаем так [6]:

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

где $height$ – высота нечеткого множества.

3. Параметрическая идентификация зависимостей с помощью базы нечетких гибридных правил

Обучающую выборку из M пар “входы – выход” запишем таким образом:

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

где X_r – входной вектор в r-ой строке выборки и y_r – соответствующий выход.

Обозначим через $y = F(P, X)$ модель на основе базы нечетких гибридных правил (1) с параметрами P. Параметрическая идентификация заключается в нахождении вектора P, который обеспечивает:

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

В (7) управляемые переменные P соответствуют параметрам функций принадлежности нечетких множеств из антецедентов и консеквентов правил (1). Для сохранения интерпретабельности модели на параметры функций принадлежности нечетких множеств наложим ограничения согласно [7].

4. Экспериментальные данные для идентификации модели прогнозирования трудоемкости

Для синтеза модели прогнозирования трудоемкости разработки программных систем воспользуемся данными, предоставленными фирмой "Орнеон". Они относятся к проекту по созданию игры типа "Quest". Выходной переменной является y – трудоемкость выполнения задачи, которую измеряют в человеко-днях. Входными переменными являются такие факторы:

x_1 – стаж программиста в месяцах;

x_2 – новизна задачи, которую оценивают в баллах от 1 до 10;

x_3 – сложность задачи, которую оценивают в баллах от 1 до 10.

Обучающая выборка сформирована из задач, которые выполнялись в течение первых трех месяцев проекта, а тестовая – из более поздних задач. В обучающую выборку попало 107 случаев (рис. 1), а в тестовую – 106 (рис. 2).

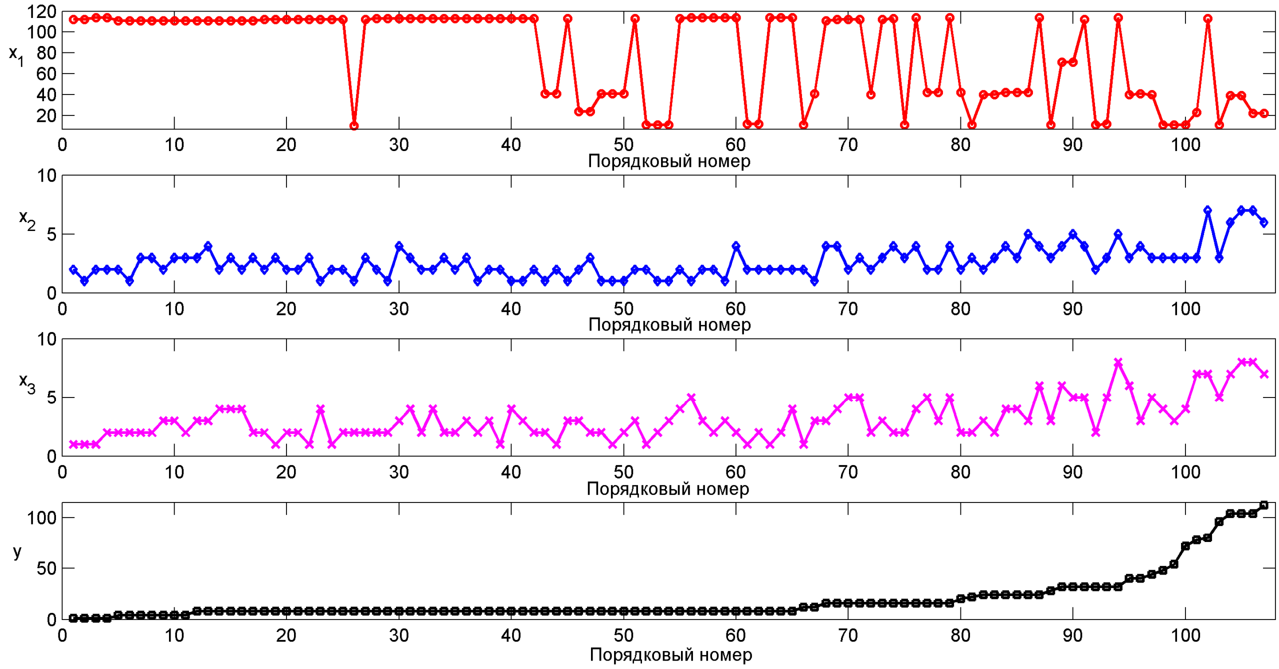

Рис. 1. Обучающая выборка

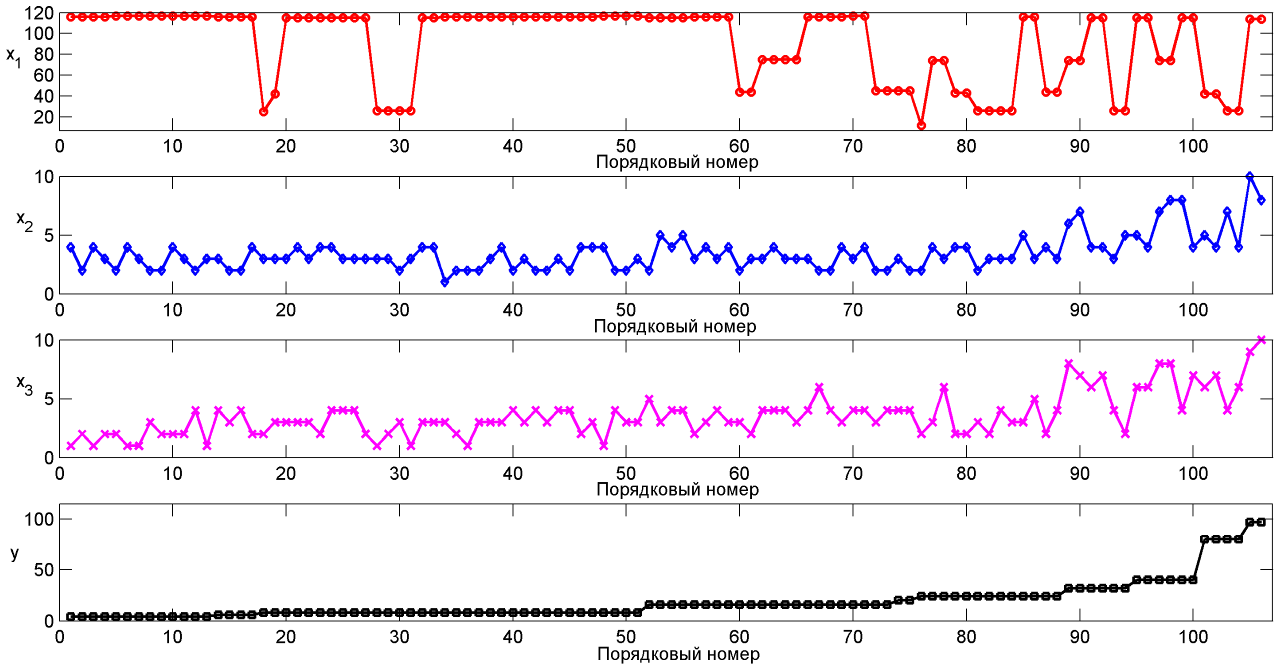

Рис. 2. Тестовая выборка

5. Модели прогнозирования трудоемкости разработки программных систем

Моделирование зависимости $y = f(x_1, x_2, x_3)$ осуществим с помощью следующих трех нечетких гибридных правил:

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

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

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

В этих правилах термы Low и $Average$ и нечеткие коэффициенты консеквентов заданы треугольными функциями принадлежности. Треугольная функция принадлежности имеет три параметра (a, b, c), которые задают носитель (a, c) и ядро (b) нечеткого множества. Терм *High* описан трапециевидной функцией принадлежности с четырьмя параметрами (a, b, c, d), которые задают носитель (a, d) и ядро (b, c) нечеткого множества. Параметры функций принадлежности термов и нечетких коэффициентов после обучения сведены в табл. 1.

Таблица 1

Параметры функций принадлежности термов и нечетких коэффициентов нечеткой гибридной модели

Нечеткое множество	Параметры функций принадлежности
$\widetilde{k}_{10}$	(-3.13, 0.87, 4.97)
$\widetilde{k}_{11}$	(-4.84, -3.26, -2.86)
$\widetilde{k}_{12}$	(7.29, 9.22, 10.71)
$\widetilde{k}_{13}$	(11.54, 12.74, 14.04)
$\widetilde{k}_{20}$	(-5.95, -4.91, -3.97)
$\widetilde{k}_{21}$	(-1.41, -1.37, -1.37)
$\widetilde{k}_{22}$	(3.97, 8.32, 10.12)
$\widetilde{k}_{23}$	(7.99, 14.97, 16.9)
$\widetilde{k}_{30}$	(-2.14, 1.88, 8.09)
$\widetilde{k}_{31}$	(-0.46, -0.46, -0.23)
$\widetilde{k}_{32}$	(9.78, 10.7, 11.55)
$\widetilde{k}_{33}$	(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)

Для сравнения качества идентификации мы синтезировали 5 конкурентных моделей:

- линейную

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

- квадратичную

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

- полином степени 1/2

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

- ряд Винера

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

- базу нечетких правил Сугено

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

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

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

Параметры функций принадлежности термов антецедентов правил Сугено сведены в табл. 2.

Таблица 2

Параметры функций принадлежности термов и нечетких коэффициентов нечеткой гибридной модели

Нечеткое множество	Параметры функций принадлежности
Low	(0, 0, 29.13)
Average	(-11.49, 28.07, 67.63)
High	(21.3, 54.8, 129.68, 138)

Сравнение результатов моделирования с экспериментальными данными приведено на рис. 3 и 4. Для некоторых случаев по конкурентным моделями спрогнозированная трудоемкость была меньше 1. Для этих случаев выходное значение установлено равным 1. Из рис. 3 и 4 видно, что как по средней квадратичной невязке ($RMSE$), так и по максимальной абсолютной невязке ($MaxErr$) нечеткая гибридная модель является наилучшей. Модель на основе правил Сугено показала близкие результаты, поскольку ее можно считать частным случаем гибридной нечеткой модели, в которой все коэффициенты в консеквентах заданы четкими числами.

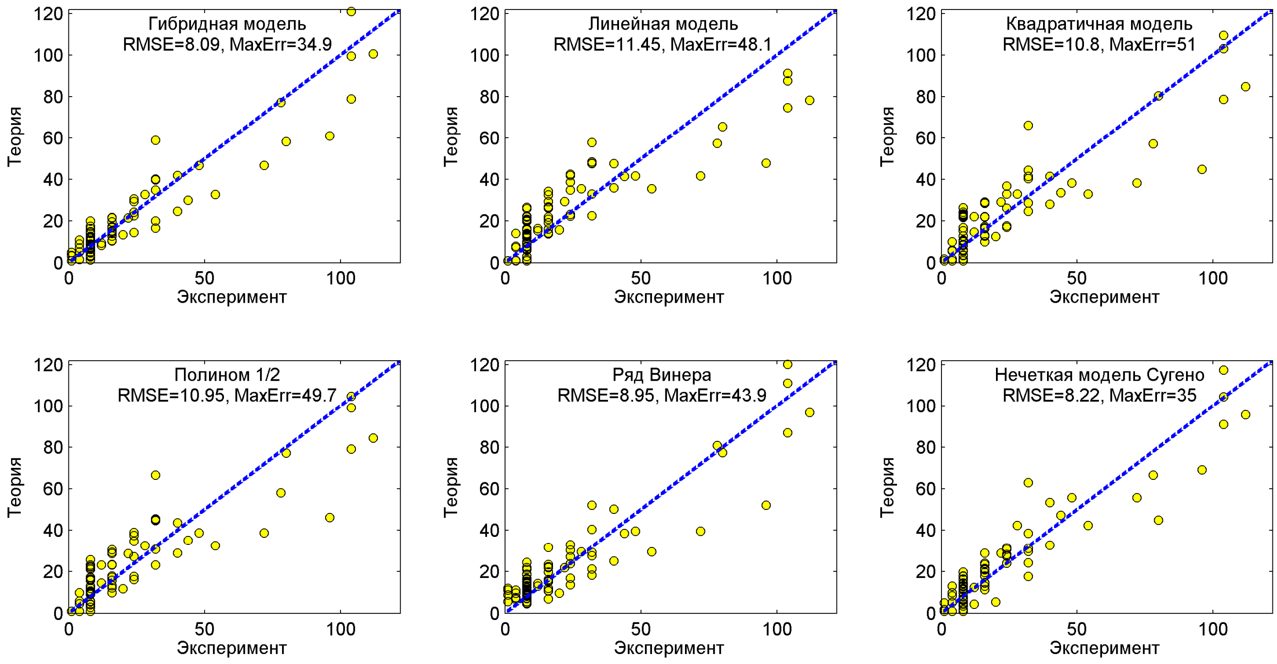

Рис. 3. Проверка моделей на обучающей выборке

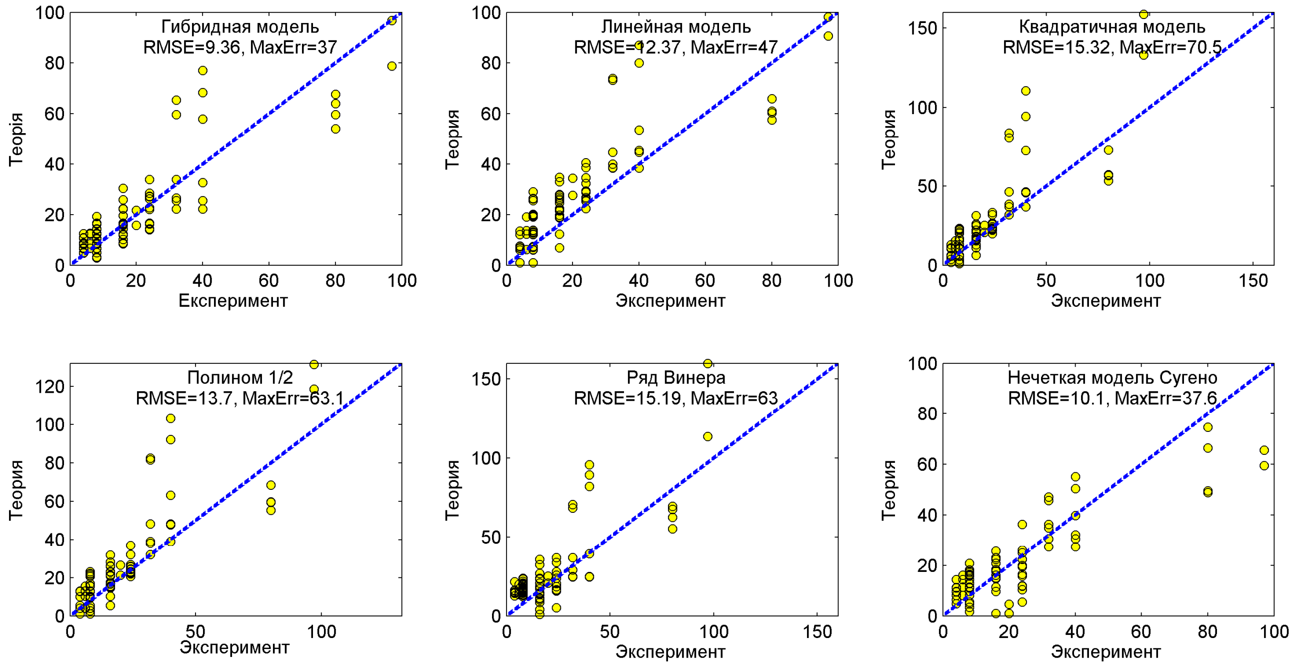

Рис. 4. Проверка моделей на тестовой выборке

Выводы

Исследовано применение нечеткой гибридной модели для прогнозирования трудоемкости разработки программных систем. Проведено сравнение результатов с альтернативными моделями. Гибридная нечеткая модель показала лучшие результаты среди рассмотренных моделей. Следовательно, ее применение для прогнозирования времени трудоемкости разработки программных систем целесообразно, поскольку это позволит улучшить проектное планирование.

СПИСОК ЛИТЕРАТУРЫ

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.

Штовба Сергей Дмитриевич – профессор, д. т. н., профессор кафедры компьютерных систем управления.

Яковенко Антон Андреевич – студент ИнАЕКСУ.

Винницкий национальный технический университет.