УДК 568.012

С.Д. Штовба

НЕЧЕТКАЯ ИДЕНТИФИКАЦИЯ НА ОСНОВЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ ПАРАМЕТРИЧЕСКОЙ ФУНКЦИИ ПРИНАДЛЕЖНОСТИ

Введение. Один из наиболее популярных методов обработки данных — регрессионный анализ. Однако он не годится для прикладных задач идентификации, в которых информация об исследуемой зависимости «входы–выход» содержит нечеткие оценки типа «низкий», «средний», «очень высокий» и т.п. В настоящей статье рассматривается задача построения нечеткой регрессионной модели по выборке данных с четкими входами и нечетким выходом. Нечеткая регрессия впервые описана в работе [1]. Она представляет собой некоторую нечеткую функцию, связывающую входы и выход исследуемой зависимости. Параметры этой функции — коэффициенты регрессии — задаются нечеткими числами. Для текущего входного вектора нечеткое значение на выходе регрессионной модели рассчитывается по принципу обобщения Заде [2].

В работе [1] задача идентификации нечетких коэффициентов регрессионной модели сведена к задаче линейного программирования. Она заключается в отыскании таких параметров функций принадлежности, которые минимизируют суммарную размазанность нечетких коэффициентов. При этом для каждого набора данных α-сечение нечеткого выхода регрессионной модели должно включать α-сечение соответствующего нечеткого числа из обучающей выборки. Выполнение этого условия необходимо обеспечить для всех α-уровней выше наперед заданного порогового значения. Основной недостаток подхода [1] заключается в высокой чувствительности коэффициентов регрессии к выбросам данным. Кроме того, целевая функция в задаче нечеткой идентификации не интерпретируется как некоторый показатель схожести желаемого и действительного поведения модели, в отличие от обычного регрессионного анализа.

В статье [3] предложено подбирать нечеткие коэффициенты регрессии таким образом, чтобы минимизировать расстояние между нечеткими числами — выходом модели и данными из обучающей выборки. Для этого применяют различные метрики [3, 4]. Соответствующая задача оптимизации становится нелинейной, поэтому для ее решения кроме градиентных методов используют и генетические алгоритмы [5].

В настоящей статье предлагается новая структура нечеткой регрессионной модели. Вместо аппроксимации зависимости «входы–выход» функцией с нечеткими коэффициентами каждой точке факторного пространства ставится в соответствие нечеткое число с параметрической функцией принадлежности. Зависимость параметров этой функции принадлежности от влияющих факторов описывается четкими моделями при помощи обычного регрессионного анализа выборки данных.

1. **Новая структура нечеткой регрессионной модели**. Рассматривается отображение вектора $X = (x_1, x_2, ..., x_n)$ четких числовых значений влияющих факторов в нечеткое значение $\tilde{y}$ функции отклика:

$$(x_1, x_2, ..., x_n) \to \tilde{y} = \int_{y \in [\underline{y}, \bar{y}]} \mu_{\tilde{y}}(y) / y,$$

где $\mu_{\tilde{y}}(y)$ — функция принадлежности нечеткого числа $\tilde{y}$ на носителе $[\underline{y}, \bar{y}]$.

Предположим, что на всем факторном пространстве искомое нечеткое число $\tilde{y}$ можно описать параметрической функцией принадлежности одного типа. Обозначим эту функцию принадлежности $\mathrm{mf}(y, Z)$, где Z — вектор параметров функции принадлежности. Зависимость $Z = f(X, P)$ опишем системой регрессионных моделей с коэффициентами P, каждая из которых связывает влияющие факторы с одним параметром функции принадлежности нечеткого числа $\tilde{y}$. Таким образом, $\mathrm{mf}(y, Z) = \mathrm{mf}(y, f(X, P))$.

Пусть, например, нечеткое число $\tilde{y}$ задано гауссовой функцией принадлежности

$$\mu(y) = \exp\left(-\frac{(y-b)^2}{2c^2}\right), \tag{1}$$

где параметры функции принадлежности b и c — координата максимума и коэффициент концентрации $(Z = (b, c))$. Тогда при использовании линейных регрессионных моделей зависимость этих параметров от факторов $X = (x_1, x_2, ..., x_n)$ записывается следующим образом:

$$b = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n,$$
$$c = c_0 + c_1 x_1 + c_2 x_2 + ... + c_n x_n,$$

где $(b_0, b_1, ..., b_n, c_0, c_1, ..., c_n) = P$ — коэффициенты регрессии.

2. **Постановка задачи нечеткого регрессионного анализа**. Нечеткую обучающую выборку определим как M пар данных

$$(X_r, \tilde{y}_r), \quad r = \overline{1, M}, \tag{2}$$

где $X_r = (x_{r1}, x_{r2}, ..., x_{rn})$ — входной вектор в r-й строке выборки, $\tilde{y}_r = \int\limits_{y \in [\underline{y}, \bar{y}]} \mu_{\tilde{y}_r}(y)/y$ — соответствующий выход в виде нечеткого числа.

Сформулируем задачу нечеткого регрессионного анализа по нечеткой выборке (2) как поиск таких коэффициентов P, которые обеспечивают

$$\sqrt{\frac{1}{M}\sum_{r=\overline{1,M}} \mathrm{RMSE}(\tilde{y}_r, \tilde{F}(P, X_r))^2} \to \min, \tag{3}$$

где $\tilde{F}(P, X_r)$ — нечеткое число с функцией принадлежности $\mathrm{mf}(y, f(X_r, P))$, полученное для входного вектора X_r по системе регрессионных моделей с коэффициентами P; RMSE — расстояние между двумя нечеткими числами, соответствующими желаемому и действительному поведению модели в точке X_r.

Расстояние между двумя нечеткими числами $\tilde{A}$ и $\tilde{B}$ с функциями принадлежности $\mu_{\tilde{A}}(y)$ и $\mu_{\tilde{B}}(y)$ на интервале $[\underline{y}, \bar{y}]$ определим так:

$$\mathrm{RMSE}(\widetilde{A}, \widetilde{B}) = \sqrt{\frac{\int_{\underline{y}}^{\overline{y}} (\mu_{\widetilde{A}}(y) - \mu_{\widetilde{B}}(y))^2 dy}{\overline{y} - \underline{y}}}. \qquad (4)$$

Формула (4) позволяет рассчитать расстояние между произвольными нечеткими числами на непрерывном носителе. Если нечеткие числа заданы на дискретном носителе $\{y_1, y_2, ..., y_K\}$, то формула (4) преобразуется к виду

$$\mathrm{RMSE}(\widetilde{A}, \widetilde{B}) = \sqrt{\frac{\sum_{j=\overline{1,K}} (\mu_{\widetilde{A}}(y_j) - \mu_{\widetilde{B}}(y_j))^2}{K}}. \qquad (5)$$

Задача (3) может быть решена методами нелинейной оптимизации,

3. Тестовая задача. В [6] приведены данные 392 экспериментов о зависимости времени y разгона автомобиля до скорости 60 миль в час от количества цилиндров x_1 и тяговооруженности автомобиля (отношения мощности к массе автомобиля) x_2. По этим данным сформируем нечеткие обучающую и тестовую выборки следующим образом.

В экспериментальных данных влияющие факторы принимают такие значения: $x_1 \in \{3, 4, 5, 6, 8\}$; $x_2 \in [0{,}0206; 0{,}729]$. Для формирования нечетких выборок округлим значения фактора x_2 до тысячных. Тогда $x_2 \in \{0{,}021; 0{,}022; \ldots$ $\ldots, 0{,}051; 0{,}054; 0{,}073\}$. Декартово произведение $x_1 \times x_2$ состоит из $5 \times 33 = 165$ точек, из них для 27 пар (x_1, x_2) в экспериментальных данных существует не менее трех различных значений выходной переменной y. Для этих 27 пар, используя идеи потенциала точки из горной кластеризации [7], рассчитаем степени принадлежности по распределению значений выходной переменной. Потенциал точки — это число, показывающее, насколько плотно в ее окрестности расположены экспериментальные данные. Чем он выше, тем ближе точка к центру кластера. Потенциал точки y_i $(i = \overline{1, v})$ рассчитывают так [7]:

$$\mathrm{pot}_i = \sum_{j=\overline{1,v}} \exp(-4\beta^2 (y_i - y_j)^2),$$

где $\beta > 0$ — коэффициент размазанности кластера; v — количество точек.

Перед использованием этой формулы спроецируем данные на единичный отрезок. Степени принадлежности нечеткого множества $\widetilde{y}$ рассчитаем из потенциалов следующим образом:

$$\mu_{\widetilde{y}}(y_i) = \frac{\mathrm{pot}_i}{\max\limits_{j=\overline{1,v}} (\mathrm{pot}_j)},,$$

Затем найденные степени принадлежности аппроксимируем двухсторонней гауссовой кривой:

$$\mu(y) = \begin{cases} \mathrm{gmf}(y, b, c_1), & \text{если } y < b, \\ \mathrm{gmf}(y, b, c_2), & \text{если } y \geq b, \end{cases}$$

где gmf — гауссова функция принадлежности (1).

В обучающую выборку включим 20 пар данных, а в тестовую — 7. В табл. 1 приведена нечеткая обучающая выборка, а в табл. 2 — нечеткая тестовая выборка.

Таблица 1

x_1	x_2	$\tilde{y}$		
		b	c_1	c_2
4	0,024	21,8	4,52	4,19
4	0,031	17,4	1,53	1,56
4	0,032	16,5	2,44	2,45
4	0,033	16,5	2,55	2,71
4	0,034	15,7	2,08	2,79
4	0,036	14,7	2,35	3,16
4	0,043	14,1	3,57	3,51
6	0,028	17,6	2,68	2,77
6	0,031	16,2	1,61	1,59
6	0,032	16	1,18	1,12
6	0,035	15,4	0,31	3,06
6	0,037	15,5	2,11	2,35
6	0,043	13,5	0,8	0,84
8	0,032	13,8	0,76	2,68
8	0,034	13,8	1,76	1,59
8	0,035	14,5	1,48	1,28
8	0,036	12,8	1,49	1,53
8	0,037	13,1	0,92	0,88
8	0,043	11,6	2,42	2,38
8	0,046	11	3,9	3,79

Таблица 2

x_1	x_2	$\tilde{y}$		
		b	c_1	c_2
4	0,028	15,3	3,42	6,93
4	0,035	15,1	3,13	3,6
4	0,037	15,3	3,35	3,89
6	0,029	16,7	1,38	1,60
6	0,034	15,9	1,14	0,86
8	0,033	14,6	1,76	1,79
8	0,044	11,2	2,26	3,23

4. Пример нечеткого регрессионного анализа. Для тестовой задачи построим линейные регрессионные модели, которые свяжут факторы x_1 и x_2 с параметрами гауссовой функции принадлежности нечеткого времени $\tilde{y}$ разгона автомобиля. В результате решения задачи (3) получаем такие модели:

$$\begin{aligned} b &= 25{,}325 - 0{,}387x_1 - 228{,}48x_2, \\ c &= 3{,}571 - 0{,}237x_1 + 0{,}009x_2. \end{aligned} \tag{6}$$

На тестовой выборке они обеспечивают невязку $\mathrm{RMSE} = 0{,}2043$. Графики функций принадлежности (рис. 1) свидетельствуют о приемлемом качестве идентификации.

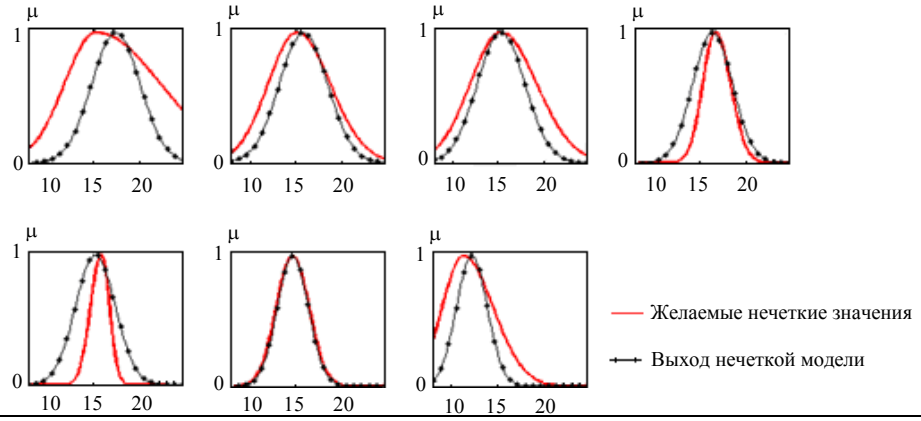

Рис. 1

Для сравнения по этим же данным построим традиционную нечеткую регрессионную модель вида

$$\tilde{y} = \tilde{a}_0 + \tilde{a}_1 x_1 + \tilde{a}_2 x_2. \qquad (7)$$

Нечеткие коэффициенты $\tilde{a}_0, \tilde{a}_1, \tilde{a}_2$ опишем двухсторонней гауссовой функцией принадлежности. Оптимальные по (3) нечеткие коэффициенты этой модели изображены на рис. 2. С этими коэффициентами невязка модели (7) на нечеткой тестовой выборке составляет $\mathrm{RMSE} = 0{,}2974$ (рис. 3), что хуже, чем в предыдущем случае.

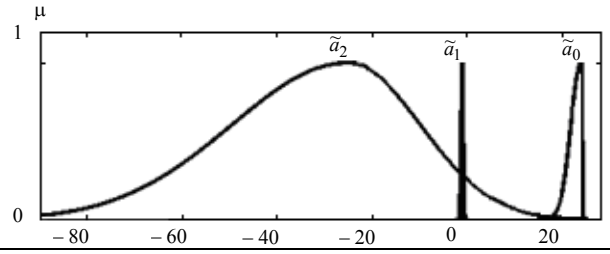

Рис. 2

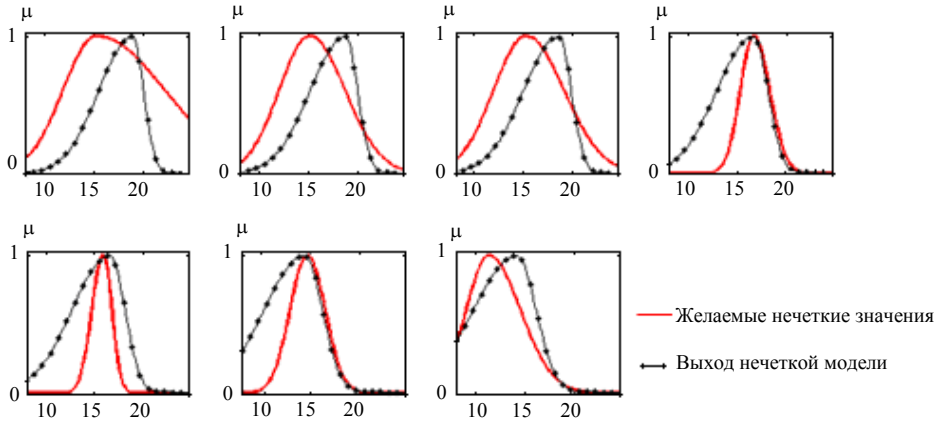

Рис. 3

5. Быстрый нечеткий линейный регрессионный анализ. Предположим, что нечеткие числа в обучающей выборке (2) заданы параметрическими функциями принадлежности одинакового типа. Обозначим параметры этой функции

принадлежности $z_1, z_2, ..., z_k$. Следовательно, обучающую выборку (2) можно записать в виде

$$((x_{r1}, x_{r2}, ..., x_{rn}), (z_{r1}, z_{r2}, ..., z_{rk})), \quad r = \overline{1, M}. \tag{8}$$

Предположим, что зависимость параметров функции принадлежности нечеткого числа $\tilde{y}$ от влияющих факторов $(x_1, x_2, ..., x_n)$ описывается линейными моделями

$$z_1 = a_{10} + a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n,$$
$$z_2 = a_{20} + a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$z_k = a_{k0} + a_{k1}x_1 + a_{k2}x_2 + ... + a_{kn}x_n.$$

Тогда отыскание коэффициентов этих моделей сводится к обычной процедуре линейного регрессионного анализа, которую по выборке (8) следует выполнить ровно *k* раз.

В качестве примера приведем построенные по данным табл. 1 линейные регрессионные зависимости параметров двухсторонней гауссовой функции принадлежности (3) от факторов x_1 и x_2:

$$b = 29{,}67 - 0{,}603x_1 - 312{,}84x_2,$$
$$c_1 = 2{,}629 - 0{,}253x_1 + 25{,}973x_2,$$
$$c_2 = 3{,}196 - 0{,}239x_1 + 15{,}61x_2.$$

Эти модели на нечеткой тестовой выборке (табл. 2) обеспечивают невязку RMSE = 0,2183.

6. Синтез нечетких множеств II типа по нечетким регрессионным моделям. Если значения влияющих факторов заданы нечеткими числами, тогда при постановке их в регрессионную модель типа (6) получаем параметры функции принадлежности в виде нечетких чисел. Применяя принцип обобщения к аналитическим формулам функций принадлежности с нечеткими параметрами, получаем нечеткие множества II типа. Например, пусть тяговооруженность автомобиля описана нечетким числом с гауссовой функцией принадлежности (1) при параметрах $b = 0{,}03$; $c = 0{,}003$ (рис. 4, *а*). Тогда для четырехцилиндрового автомобиля $(x_1 = 4)$ по формулам (6) получаем такие нечеткие значения параметров гауссовой функции принадлежности времени разгона автомобиля (рис. 4, *б*):

$$\tilde{c} = (2{,}6232; 2{,}6234)_{0{,}01} \cup (2{,}6232; 2{,}6233)_{0{,}5} \cup (2{,}6233; 2{,}6233)_1;$$

$$\tilde{b} = (15{,}45; 18{,}39)_{0{,}01} \cup (16{,}12; 17{,}73)_{0{,}5} \cup (16{,}92; 16{,}92)_1.$$

Подставляя найденные нечеткие параметры в формулу (6) по α-уровневому принципу обобщения получаем значение времени разгона в виде нечеткого числа II типа. На рис. 4, *в* оно изображено совокупностью α-срезов степеней принадлежности.

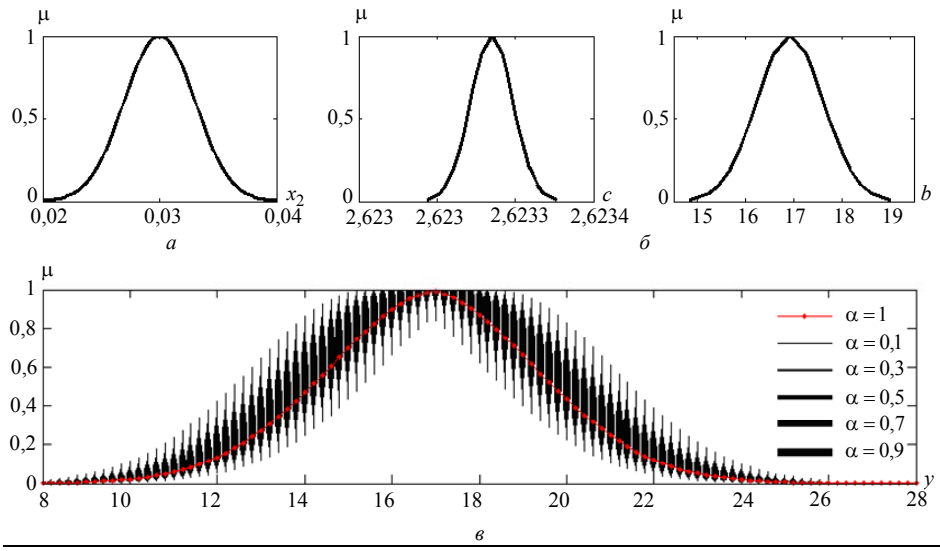

Рис. 4

Выводы. Предложена новая структура нечеткой регрессионной модели, по которой каждой точке факторного пространства ставится в соответствие нечеткое число с параметрической функцией принадлежности. Зависимость параметров этой функции принадлежности от влияющих факторов описывается четкими регрессионными моделями. Коэффициенты регрессии определяются путем минимизации суммарного расстояния между нечеткими числами — результатами моделирования и значением функции отклика в обучающей выборке.

Предложенный подход упрощает процедуру регрессионного анализа выборок данных с нечеткими значениями. При этом, как показывают компьютерные эксперименты, точность аппроксимации не ухудшается. Кроме того, согласно предложенной модели функция отклика рассчитывается по четким числам без применения принципа обобщения, что значительно сокращает вычислительную сложность по сравнению с традиционной нечеткой регрессионной моделью. При подстановке в предложенные регрессионные модели нечетких значений факторов функция отклика получается в виде нечеткого множества II типа.

Показано, что для нечетких обучающих выборок с выходными данными в виде параметрических функций принадлежности одного типа задача синтеза нечеткой модели «входы–выход» сводится к обычному многофакторному регрессионному анализу. Такой регрессионный анализ необходимо выполнять отдельно для каждого параметра функции принадлежности. Функции принадлежности обычно имеют 2–4 параметра, поэтому вычислительная сложность нечеткого регрессионного анализа будет лишь в 2–4 раза выше, чем для четкого.

Описанные преимущества предложенной структуры нечеткой регрессионной модели позволяют ей конкурировать с другими методами обработки нечетких выборок данных в инженерии, медицине, политике, социологии, политологии, спорте и в других областях.

С.Д. Штовба

НЕЧІТКА ІДЕНТИФІКАЦІЯ НА ОСНОВІ РЕГРЕСІЙНИХ МОДЕЛЕЙ ПАРАМЕТРИЧНОЇ ФУНКЦІЇ НАЛЕЖНОСТІ

Запропоновано нову структуру нечіткої регресійної моделі, за якої кожній точці факторного простору ставиться у відповідність нечітке число з параметричною функцією належності. Залежність параметрів цієї функції належності від факторів впливу описується чіткими регресійними моделями. Коефіцієнти регресії визначаються за нечіткою навчальною вибіркою.

S.D. Shtovba

FUZZY IDENTIFICATION BASED ON REGRESION MODELS OF PARAMETRICAL MEMBERSHIP FUNCTION

A new structure of fuzzy regression model is proposed. The model maps an input vector into output fuzzy number with parametrical membership function. Crisp regression models take into account a dependence of the membership function parameters upon the influence factors. The regression coefficients are calculated based on learning sample.

1. *Tanaka H., Uejima S., Asai K.* Linear regression analysis with fuzzy model // IEEE Trans. Systems Man Cybernet. — 1982. — **12**, N 6. — P. 903–907.
2. *Zadeh L.* Fuzzy sets // Information and Control. — 1965. — N 8. — P. 338–353.
3. *Diamond P.* Fuzzy least squares // Information Sci. — 1988. — **46**, N 3. — P. 141–157.
4. *Papadopoulos B., Sirpi M.* Similarities and distances in fuzzy regression modeling // Soft Computing. — 2004. — **8**, N 8. — P. 556–561.
5. *Aliev R., Fazlollahi B., Vahidov R.* Genetic algorithms-based fuzzy regression analysis // Soft Computing. — 2002. — **6**, N 6. — P. 470–475.
6. *MPG* data base of UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html).
7. *Yager R., Filev D.* Essentials of fuzzy modeling and control. — : John Wiley & Sons, 1994. — 387 p.

Получено 28.07.2006

 ISSN 0572-2691