КОМПЬЮТЕРНАЯ МАТЕМАТИКА

УДК 519.254

С.В. ЄРШОВ, Т.І. ЛИКО

МЕТОДИ ПОБУДОВИ РЕГРЕСІЙНИХ МОДЕЛЕЙ НА ОСНОВІ НЕЧІТКИХ ДАНИХ

Запропоновано метод побудови регресійних моделей для систем на основі нечітких правил у випадку, коли реакція систем представлена нечіткими даними. Розроблено алгоритм, який з прийнятною точністю будує адекватну кількість правил Такагі-Сугено регресійної моделі з використанням автоматичної стратегії на основі даних спостережень, що надходять. Побудовано процедуру, що використовується для знаходження максимальної схожості параметрів регресійних моделей, у випадку, коли модель залежить від параметрів у консеквентах нечітких правил.

В даний час велика увага приділяється методам представлення моделей складних систем і процесів за допомогою нечіткої математики. Такі моделі на сьогодні активно використовуються в бізнесі, управлінні, економіці, для вирішення задач розробки програмного забезпечення, а також у багатьох інших сферах, таких як охорона здоров’я та біологія. Однак, застосуванню нечіткої логіки для побудови кореляцій між залежними і незалежними змінними процесу, тобто регресійного моделювання, приділялося значно менше уваги.

Нечіткі моделі, що задаються нечіткими правилами Такагі-Сугно або Мамдані [1], є ефективним засобом, який використовується для апроксимації поведінки нелінійних систем та процесів. Ситуація ускладнюється через те, що у багатьох реальних моделях, має бути врахована нечіткість даних, на основі яких вони побудовані. Така нечіткість може виникати через обмежену точність вимірювальних пристроїв, неповноту і недовизначеність опису об’єктів даних.

Існує дві основні проблеми, які заважають побудувати нечіткі регресійні моделі на основі правил у випадку нечітких даних. Перша проблема – яким чином визначити нечіткі множини антецедентів правил, що ґрунтуються на нечітких даних. Друга проблема, яка вирішена в даній роботі – в тому, як визначити лінійні функції у консеквентах правил, якщо спостереження виходу представлені нечіткими даними.

Мета даної роботи – побудова систематичного методу для розробки нечітких регресійних моделей, що ґрунтуються на правилах, у випадку нечітких даних. Дана робота являє собою подальший розвиток досліджень у напрямку становлення адаптивних моделей та архітектур інтелектуальних мультиагентних систем у нечіткому представленні [2 – 4].

Нечіткий регресійний аналіз є нечітким типом класичного регресійого аналізу, використовується для оцінки функціональних відношень між залежними і незалежними змінними в нечіткому середовищі. У всіх випадках нечіткої регресії, лінійна регресія рекомендується для практичних ситуацій, коли рішення часто мають бути зроблені на основі неточних, неповних даних. Існує багато підходів для вирішення таких задач. Модель нечіткої регресії вперше розроблена Танакою [5] у вигляді лінійної системи. Танака та інші спочатку застосовували нечітку процедуру лінійної регресії для чітких експериментальних даних. У подальших роботах використовувались нечіткі дані на вході для побудови нечіткої регресійної моделі. Нечіткі вхідні дані, що використовувались у цих експериментах, представлені у вигляді трикутних нечітких чисел [6]. Метод найменших квадратів для нечітких змінних, що запропонований Даймондом [7], дає рішення аналога нормального рівняння класичних найменших квадратів.

Задачею регресійного аналізу є знаходження функціональних відношень між незалежними змінними $x = [x_1, x_2, ..., x_n]$ та залежними змінними $y = [y_1, y_2, ..., y_n]$, де n – кількість незалежних змінних, а m – кількість залежних змінних. Регресійна модель виражається в цьому випадку так:

$$y = f(x) + \varepsilon, \tag{1}$$

де $f(x) = [f_1(x), f_2(x), ..., f_m(x)]$, а $\varepsilon = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_m]$ – вектор випадкових помилок функціонального наближення. Для спрощення такої моделі припускається, що залежність між залежними і незалежними змінними є лінійною. Таким чином задача спрощується до знаходження параметрів лінійної функції. Окрім того для кожної залежної змінної будується окрема регресійна модель. В найпростішому випадку для одної незалежної змінної:

$$y = a_0 + a_1 x + \varepsilon. \tag{2}$$

Для багатьох змінних:

$$y = a_0 + a_1 x_1 + ... + a_m x_m + \varepsilon. \tag{3}$$

Для узагальнення цієї моделі в нечіткому середовищі використовується сімейство нечітких множин, що називається нечіткими числами [1].

Нечітка підмножина A множини цілих чисел R з функцією належності $\mu : \mathrm{R} \to [0,1]$ називається нечітким числом, якщо

(i) A – нормальна, тобто існує такий елемент a_0, що $\mu_\mathrm{A}(\mathrm{a}_0) = 1$;

(ii) A – нечітка випукла множина, тобто $\forall a_1, a_2 \in R \mu_A(\lambda a_1 + (1-\lambda) a_2) \geq$ $\geq \mu_A(a_1) \wedge \mu_A(a_2), \forall \lambda \in [0,1]$;

(iii) μ_A – кусково-неперервна функція;

(iv) $\sup(A) = \{a \in R : \mu_A(a) > 0\}$ є обмеженим.

Нечітке число може бути представлене сімейством множин, що називаються α -зрізами:

$$A_\alpha = \{a \in R : \mu_A(a) \geq \alpha\}. \quad (4)$$

Тоді нечітке число *A* можна представити, як об'єднання множин α -зрізами:

$$A = \bigcup_{\alpha \in [0,1]} A_\alpha.$$

З визначення нечіткого числа легко бачити, що кожен α -зріз нечіткого числа *A* – закритий інтервал $A_\alpha = [A^L(\alpha), A^U(\alpha)],$ де

$$A^L(\alpha) = \inf\{a \in R : \mu_A(a) \geq \alpha\}, \quad (5)$$

$$A^U(\alpha) = \sup\{a \in R : \mu_A(a) \geq \alpha\}. \quad (6)$$

Звідси, для двох нечітких чисел *A* і *B* з α -зрізами $A_\alpha = [A^U(\alpha), A^U(\alpha)],$ $B_\alpha = [B^U(\alpha), B^U(\alpha)],$ можна визначити відстань між *A* і *B*:

$$d(A,B) = \sqrt{\int_0^1 (A^L(\alpha) - B^L(\alpha))^2 da + \int_0^1 (A^U(\alpha) - B^U(\alpha))^2 da}. \quad (7)$$

Тоді для простої нечіткої регресії $Y = b_0 + b_1 X$ параметри $b_0, b_1 \in R$ можна оцінити, мінімізуючи відстань між спостережуваними даними та даними, отриманими з оцінки моделі.

$$\min H(b_0, b_1) = \sum_{i=1}^{k} d^2(Y_i, b_0 + b_1 X_i). \quad (8)$$

Знайти розв'язок такої задачі для простої нечіткої регресії можна аналітично. Однак, варто відмітити, що вигляд функції $H()$ залежить від знаку параметра b_1.

$$H^+(b_0, b_1) = \sum_{i=1}^{k} \int_0^1 (Y^L{}_i(\alpha) - b_0 - b_1 X^L{}_i(\alpha))^2 d\alpha + \int_0^1 (Y^U{}_i(\alpha) - b_0 - b_1 X^U{}_i(\alpha))^2 d\alpha. \quad (9)$$

$$H^-(b_0, b_1) = \sum_{i=1}^{k} \int_0^1 (Y^L{}_i(\alpha) - b_0 - b_1 X^U{}_i(\alpha))^2 d\alpha + \int_0^1 (Y^U{}_i(\alpha) - b_0 - b_1 X^L{}_i(\alpha))^2 d\alpha. \quad (10)$$

Рівність (9) дійсна для $b_1 > 0,$ а (10) – в іншому випадку.

З цих нерівностей можна легко знайти розв'язки для випадку однієї незалежної змінної, однак для багатьох змінних складність задачі виростає експоненціально. Однак на основі цих обчислень можна адаптувати метод градієнтного спуску для ітеративного вирішення задачі, що значно збільшить ефективність для регресії з багатьма змінними.

Алгоритм методу градієнтного спуску для однієї змінної містить такі кроки:

1) ініціалізація початкових значень моделі b_0^0, b_1^0;

2) початок ітерації $i = 1$;

3) оцінити функцію H з відповідними параметрами;

4) оновити параметри $\Delta b_0 = \mu_0 \dfrac{\partial H^+(b_0,b_1)}{\partial b_0}$ і $\Delta b_1 = \mu_0 \dfrac{\partial H^+(b_0,b_1)}{\partial b_1}$ або

$$\Delta b_0 = \mu_0 \frac{\partial H^-(b_0,b_1)}{\partial b_0} \quad \text{і} \quad \Delta b_1 = \mu_0 \frac{\partial H^-(b_0,b_1)}{\partial b_1};$$

5) оновити оцінку параметрів $b_0^{\,i} = b_0^{\,i-1} + \Delta b_0$ та $b_1^{\,i} = b_1^{\,i-1} + \Delta b_1$;

6) якщо $\Delta b_0 > \varepsilon$ або $\Delta b_1 > \varepsilon$, то $i = i+1$ і продовжити з п. 3, інакше зупинка.

Нехай $u = [u_1, u_2, ..., u_p]^T$ – вектор вхідних змінних, а x – вихідна змінна. Нечітка модель (типу Такагі-Сугено) складається з набору нечітких правил IF-THEN, що мають такий вигляд:

$$R_k\text{: } IF\ u_1 = A_{k1}\ I \ldots I\ u_p = A_{kp}\ \text{THEN}\ x^k = b_{k0} + b_{k1}u_1 + b_{k2}u_2 + ... + b_{kp}u_p,$$

де k = 1,…, M, x^k – це вихід k-го правила, $b_k = [b_{k1}, b_{k2}, ..., b_{kp}]^T$ – коефіцієнти в k-му правилі, A_{kj} (j = 1,2,..., p) – нечіткі множини антецедента правила. Враховуючи довільне введення даних $u_i = [u_{i1}, u_{i2}, ..., u_{ip}]^T$, кожне правило забезпечує прогнозований вихід. Агрегований вихід нечіткої моделі Такагі-Сугено обчислюється таким чином:

$$x_i = \sum\nolimits_{k=1}^{M} \tau_i^k x_i^k \Big/ \sum\nolimits_{k=1}^{M} \tau_i^k x, \tag{11}$$

де τ_i^k – сила «спрацювання» правила R_k для i-го входу.

Розглянемо стратегію ідентифікації антецеденту нечіткої моделі, при якому спостереження в наступному вигляді:

$$T = \{e_i \,|\, e_i = (u_i, \breve{x}_i), i = 1,...,n\},$$

де $\widehat{x}_i$ – неточно спостережувані значення для виходу x.

Початкове нечітке правило генерується за допомогою простого способу. Нечітких множини A_{kj} в антецеденті початкового нечіткого правила визначаються наступним чином:

$$v_{1j} = \frac{1}{n}\sum_{i=1}^{n} u_{ij}, \quad s_{1j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (u_{ij} - v_{1j})^2},$$

де v_{kj} і s_{kj} (κ = 1,…, M, j = 1,…, p) представляють відповідно центр і ширину в правилі. При цьому консеквент правила ідентифікується з використанням алгоритму, який описаний далі.

Будується нове правило і додається до вихідної бази правил. Вектор, що має найгірше значення M_f, позначається u_i та розглядається як центр цього нового правила:

$$M_f = \frac{1}{n}\sum_{i=1}^{n}\left|E(\hat{x}_i) - x_i\right|^2,$$

де $E(\hat{x}_i) = \int_{\Omega_X} \omega\mu_{\hat{x}_i}(\omega)\,d\omega$ – очікувана схожість, що пов'язана з $\hat{x}_i$.

Нечіткі антецеденти $A_{нов,\,j}$ нового згенерованого правила будуть характеризуватися значеннями $v_{нов,\,j}$, $s_{нов,\,j}$, які визначаються наступним чином:

$$v_{нов,\,j} = u_{i,\,j},\ j = 1,...,\ p,$$

$$s_{нов,\,j} = \sqrt{\sum\nolimits_{i=1}^{n}\mu_{i,i}^2 \left|u_{i,j} - v_{нов,\,j}\right|^2 / \sum\nolimits_{i=1}^{n}\mu_{i,i}^2}\,.$$

Нарешті, коли нове нечітке правило додається до бази правил, номер правила, збільшується, тобто $M = M + 1$, і маємо $v_{M,\,j} = v_{нов,\,j}$, $s_{M,\,j} = s_{нов,\,j}$.

Перетворимо узагальнений консеквент (11) у наступний векторний або матричний вигляд $x_i = h_i^T b, i = 1,..., n,$ або $x = H^T b,$ де h, H і B визначаються відповідно наступним чином:

$$h_i = [w_i^1, w_i^1 u_{i1},..., w_i^1 u_{ip} ..., w_i^M, w_i^M u_{i1},..., w_i^M u_{ip}],$$

$$H = [h_1, h_2,..., h_n],\ \ b = [b_1^T, b_2^T,..., b_n^T]^T,$$

де w_i^k – це вага i-го значення даних в k-му правилі, що визначається як $w_i^k = \tau_i^k / \sum\nolimits_{j=1}^{M}\tau_i^j$.

Можна припустити, що кожен компонент x_i повного вектора даних x – реалізація нормальної випадкової змінної X_i із середнім $h_i^T b$ і стандартним відхиленням σ. При такому припущенні, повний вектор параметрів містить компоненти $\psi = (b^T, \sigma)$, які мають бути визначені у випадку, коли тільки x_i можна спостерігати.

Функція щільності ймовірності від X позначається $g(x, \psi)$, де $\psi = [\psi_1, \psi_2,..., \psi_d]^T$ є вектор-стовпець невідомих параметрів у просторі параметрів Ω_ψ. Алгоритмом підходить до проблеми максимізації ймовірності даних, що спостерігаються, $L(\psi, x)$, обробляючи ітераційно ймовірність повних даних $L(\psi, x) = g(\psi, x)$. Кожна ітерація алгоритму складається з двох етапів, що називаються крок визначення очікування і крок максимізації. Крок визначення полягає в обчисленні

$$Q(\psi, \psi^{(q)}) = \frac{\int \mu_{\hat{x}} \log(L(\psi, x))\ g(x, \psi^{(q)})\ dx}{L(\psi^{(q)}, \hat{x})},$$

де очікування $\log L(\psi, x)$ береться щодо умовної щільності x при $\hat{x}$, використовуючи вектор $\psi^{(q)}$. Другий крок вимагає максимізації $Q(\psi, \psi^{(q)})$ щодо ψ за простором параметрів Ω_ψ. Алгоритм поперемінно повторює вищевказані кроки, поки збільшення ймовірності даних, що спостерігаються, не стає менше деякого порога.

При цьому функція щільності ймовірності повних даних задана як

$$g(x,\psi)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x_i-h_i^T b)^2}{2\sigma^2}\right).$$

Необхідно знайти максимум $Q(\psi,\psi^{(q)})$ по ψ. Це може бути досягнуто шляхом диференціювання $Q(\psi,\psi^{(q)})$ по b і σ, відповідно, що призводить до:

$$\frac{\partial Q(\psi,\psi^{(q)})}{\partial b}=-\frac{1}{\sigma^2}\left(-H\beta^{(q)}+HH^T b\right),$$

$$\frac{\partial Q(\psi,\psi^{(q)})}{\partial \sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\left(\sum_{i=1}^{n}\alpha_i^{(q)}-2b^T \mathrm{H}\beta^{(q)}+b^T HH^T b\right).$$

Прирівнюючи ці похідні нулю, отримуємо наступні значення для b і σ:

$$b^{(q+1)}=(H\cdot H^T)^{-1}H\beta,\quad \sigma^{(q+1)}=\sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}\alpha_i^{(q)}-2b^{T(q+1)}\mathrm{H}\beta^{(q)}+b^{T(q+1)}HH^T b^{(q+1)}\right)}.$$

Коли ітерації закінчуються, можна отримати коефіцієнти регресії b і, таким чином, отримати моделі множинної регресії з чіткими входами і нечіткими виходами.

Висновки. Таким чином, у даній роботі запропоновано метод побудови регресійних моделей для систем на основі нечітких правил, коли реакція систем представлена нечіткими даними. Розроблено алгоритм, який з прийнятною точністю будує адекватну кількість правил Такагі-Сугено регресійної моделі з використанням автоматичної стратегії на основі даних спостережень, що надходять. Побудовано процедуру, що використовується для знаходження максимальної схожості параметрів регресійних моделей, у випадку, коли модель залежить від параметрів у консеквентах нечітких правил.

С.В. Ершов, Т.И. Лыко

МЕТОДЫ ПОСТРОЕНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ
НА ОСНОВЕ НЕЧЕТКИХ ДАННЫХ

Предложен метод построения регрессионных моделей для систем на основе нечетких правил, в ситуации, когда реакция систем представлена нечеткими данными. Разработан алгоритм, который с приемлемой точностью строит адекватное количество правил Такаги-Сугено регрессионной модели с использованием автоматической стратегии на основе поступающих данных наблюдений. Построена процедура, которая используется для нахождения максимального сходства параметров регрессионных моделей, в случае, когда модель зависит от параметров в консеквентах нечетких правил.

S.V. Yershov, T.I. Lyko

METHODS FOR CONSTRUCTION OF REGRESSION MODELS BASED ON FUZZY DATA

A method for construction of regression models for systems based on fuzzy rules in situation, when reaction of a system is presented by fuzzy data, is proposed. An algorithm, which builds an adequate amount of Takagi-Sugeno rules for regression model with a reasonable accuracy and uses an automated strategy based on incoming data of observations, is developed. A procedure used for finding the maximum parameter similarity of regression models when the model depends on parameters in consequents of fuzzy rules, is constructed.

1. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетические алгоритмы и нечеткие системы. – М.: Горячая линия – Телеком, 2004. – 452 с.
2. *Ершов С.В.* Модель интеллектуальных агентов, основанная на нечеткой логике высшего типа // Компьютерная математика. – 2012. – № 1. – С. 10 – 16.
3. *Ершов С.В.* Трансформационный подход к разработке адаптивных интеллектуальных агентов на основе нечетких схем переходов // Там же. – 2011. – № 1. – С. 69 – 78.
4. *Ершов С.В.* Принципы построения нечетких мультиагентных систем в распределенной среде // Там же. – 2009. – № 2. – С. 54 – 61.
5. *Tanaka H., Uejima S., Asia K.* Fuzzy linear regression analysis with fuzzy model // IEEE Transactions on Systems, Man and Cybernetics. – 1982. – N 12. – P. 903 – 907.
6. *Dubois D., Prade H.* Fuzzy Sets and Systems: Theory and Applications. – New York: Academic Press, 1980. – 392 p.
7. *Diamond P.* Fuzzy least squares // Information Sciences. – 1988. – N 46. – P. 141 – 157.

Одержано 27.01.2015

Об авторах:

Єршов Сергій Володимирович,
доктор фізико-математичних наук, завідувач відділу
Інституту кібернетики імені В.М. Глушкова НАН України,
E-mail: sershv@ukr.net

Лико Тетяна Іванівна,
аспірантка Інституту кібернетики імені В.М. Глушкова НАН України.
E-mail: lykotan@gmail.com