

УДК 004.852

© 2014 г. **И.А. Ходашинский**, д-р техн. наук

(Томский государственный университет систем управления и радиоэлектроники)

ПОСТРОЕНИЕ КОМПАКТНЫХ И ТОЧНЫХ НЕЧЕТКИХ МОДЕЛЕЙ НА ОСНОВЕ СТАТИСТИЧЕСКИХ ИНФОРМАЦИОННЫХ КРИТЕРИЕВ*

Излагается выбор нечеткой модели из множества моделей на основе трех информационных критериев: Акаике, Байеса и Хеннана-Куинна. Для оптимизации параметров antecedentов используются алгоритм дифференциальной эволюции и алгоритмы муравьиной колонии, для оптимизации параметров consequentов – метод наименьших квадратов. Приведены результаты экспериментов на идеальных и реальных данных.

Ключевые слова: нечеткая модель, алгоритмы муравьиной колонии, алгоритм дифференциальной эволюции, критерий Акаике, критерий Байеса, критерий Хеннана-Куинна.

Введение

Нечеткое моделирование – это технология, используемая для описания особенностей решаемых задач с помощью нечетких правил вывода типа ЕСЛИ – ТО.

Популярность и практичность нечетких моделей объясняется следующими причинами: природа нечетких правил, являющихся основой нечетких моделей, позволяет описать поведение моделируемой системы в терминах причинно-следственных отношений; системы, построенные на основе нечетких моделей, являются универсальными аппроксиматорами, способными представить любую непрерывную нелинейную функцию с наперед заданной степенью точности [1].

Основные принципы построения нечетких моделей: 1) база правил не должна быть большой, правила должны иметь достаточную обобщающую способность; 2) правила не обязательно должны включать в себя все входные переменные, допустимы так называемые неполные правила; 3) нечеткие термы, описывающие входные переменные, должны покрывать всю область определения переменных [2].

К нечеткой модели, построенной на основе реальных данных, выдвигаются два основных требования: 1) она должна точно воспроизводить данные из анализируемой таблицы наблюдений; 2) нечеткие правила могут быть интерпретирова-

* Работа выполнена при финансовой поддержке РФФИ (проект № 12-07-00055), РГНФ (проект № 12-06-12008) и в соответствии с Госзаданием 7.701.2011.

ны пользователем в контексте данного приложения. Интерпретируемость может быть оценена либо как сложность, выраженная через число правил, переменных, нечетких термов, либо через семантическую целостность, подразумевающую придание смысла функциям принадлежности. Обе названных цели – точность и интерпретируемость – являются противоречивыми и должны быть учтены при построении нечеткой модели [3 – 7].

В нашей работе интерпретируемость оценивается через сложность, а нечеткие модели, учитывающие компромисс между точностью и сложностью, предлагается строить на основе статистических информационных критериев.

Постановка задачи

Нечеткая модель задается правилами следующего вида:

$$\text{IF } x_1=A_{1i} \text{ AND } x_2=A_{2i} \text{ AND } \dots \text{ AND } x_n=A_{ni} \text{ THEN } y = r_i, \quad (1)$$

где A_{ji} – лингвистический терм, которым оценивается входная переменная x_j ; r_i – действительное число, которым оценивается выход y .

Нечеткая модель осуществляет отображение следующим образом:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{\sum_{i=1}^R \mu_{A_{1i}}(x_1) \cdot \mu_{A_{2i}}(x_2) \cdot \dots \cdot \mu_{A_{ni}}(x_n) \cdot r_i}{\sum_{i=1}^R \mu_{A_{1i}}(x_1) \cdot \mu_{A_{2i}}(x_2) \cdot \dots \cdot \mu_{A_{ni}}(x_n)},$$

где \mathbf{x} – входной вектор; R – число правил; n – количество входных переменных; $\mu_{A_{ij}}$ – функция принадлежности j -й входной переменной; $\boldsymbol{\theta} = \|\theta_1, \dots, \theta_M\|$ – вектор параметров нечеткой модели.

Пусть имеется таблица наблюдений $\{(\mathbf{x}_p, t_p), p = 1, \dots, m\}$, тогда среднеквадратические функции ошибки:

$$MSE(\boldsymbol{\theta}) = \sum_{p=1}^m t_p - f(\mathbf{x}_p, \boldsymbol{\theta})^2 / m. \quad (2)$$

Сложность нечеткой модели определяется количеством antecedentных и консеквентных параметров в правилах, а также количеством самих нечетких правил в модели [4 – 8]. В данной работе сложность определяется как сумма количества правил и количества нечетких термов в модели.

Ошибка MSE уменьшается по мере усложнения модели, т.е. увеличения числа правил и нечетких термов. Для соблюдения компромисса между сложностью и точностью модели будем использовать три статистических информационных критерия: критерий AIC (Akaike Information Criterion) [8], BIC (Bayesian Information Criterion) [9], HQC (Hannan-Quinn Information Criterion) [10]. Алгебраически критерии выражаются через сумму меры ошибки модели и штрафа за число параметров модели. Информационный характер критериев связан с концепцией информации Кульбака-Лейблера. Различия между критериями заключаются в определении штрафа (второе слагаемое). С учетом специфики нечеткого моделирования k -я сгенерированная модель будет иметь следующие оценки:

$$AIC(k) = \ln MSE(\boldsymbol{\theta}) + \frac{2}{m} (m_a + 1 + cR),$$

$$BIC(k) = \ln MSE(\theta) + \frac{\ln m}{m} (m_a + 1 + cR),$$

$$HQC(k) = \ln MSE(\theta) + \frac{2 \ln(\ln m)}{m} (m_a + 1 + cR),$$

где m – количество наблюдений; m_a – общее число параметров antecedентов; R – число правил; c – эмпирический коэффициент, учитывающий относительную стоимость каждого правила [8].

Согласно приведенным выше критериям, из множества сгенерированных моделей выбирается та, у которой значение оценки будет минимальным. Таким образом достигается компромисс между точностью и сложностью системы.

Генерация структуры и оптимизация параметров

В нашей работе на этапе генерации структуры задается сложность модели, определяемая как сумма числа нечетких термов и нечетких правил. Для генерации структур предлагается использовать алгоритм генерации базы правил нечеткой модели равномерным разбиением и перебором. Здесь генерируется заданное количество термов на каждую входную переменную, покрывающие всю область определения переменной. Из полученных термов путем полного перебора сочетаний термов по разным входным переменным формируются antecedенты правил, консеквенты формируются методом ближайшего соседа.

Для оптимизации параметров antecedентов правил в работе используются метод дифференциальной эволюции и непрерывный и прямой алгоритмы муравьиной колонии, а для оптимизации параметров консеквентов правил применяется метод наименьших квадратов.

Метод дифференциальной эволюции. Метаэвристика под названием «дифференциальная эволюция» была разработана Р. Сторном и К. Прайсом, это метод многомерной математической оптимизации, относящийся к классу стохастических алгоритмов [11].

В методе дифференциальной эволюции (ДЭ) генерируется некоторое множество хромосом или векторов параметров, каждый из которых содержит, помимо значений параметров нечеткой системы, еще и ошибку, вычисленную по формуле (2). Множество векторов-хромосом называют поколением. На каждой итерации порождается новое поколение, полученное из предыдущего применением специальной процедуры, объединяющей в себе операции мутации и кроссовера. Число векторов в каждом поколении неизменно и является параметром метода.

Новое поколение генерируется следующим образом. В предыдущем поколении определяется вектор с минимальной ошибкой θ_{best} . Из предыдущего поколения для каждого вектора θ_k выбираются четыре случайных вектора θ_{s1} , θ_{s2} , θ_{s3} , θ_{s4} , и генерируется мутантный вектор следующим образом:

$$\theta_{mv} = (\theta_{s1} - \theta_{s2}) + (\theta_{s3} - \theta_{s4}).$$

На основе лучшего и мутантного вектора генерируется новый вектор по следующей формуле:

$$\theta_{nb} = \theta_{best} + F * \theta_{mv},$$

где F – действительное число в интервале $[0, 1]$.

Над полученным вектором θ_{nb} выполняется специальная операция кроссовера, в результате получается вектор θ_{new} . Если для этого вектора ошибка нечеткой системы меньше, чем ошибка для вектора θ_k , то в новом поколении вектор θ_k заменяется вектором θ_{new} , иначе в новое поколение перейдет вектор θ_k .

Условием окончания работы алгоритма является достижение определенного числа итераций либо получение ошибки, меньше заданной [12].

Непрерывный алгоритм муравьиной колонии [13, 14]. Алгоритм предложен К. Socha и М. Dorigo, для выбора пути здесь используется функция плотности вероятности с гауссовым ядром [15]. Гауссово ядро $G^i(x)$ в работе основано на взвешенной сумме нескольких одномерных гауссовых функций g_l^i :

$$G^i(x) = \sum_{l=1}^k \omega_l g_l^i(x) = \sum_{l=1}^k \omega_l \frac{1}{\sigma_l^i \sqrt{2\pi}} e^{-\left(\frac{x-\theta_l^i}{\sqrt{2}\sigma_l^i}\right)^2}.$$

Каждому параметру нечеткой модели соответствует свое гауссово ядро, $i = 1, \dots, N$; N – число параметров. Каждая функция $G^i(x)$ описывается тремя векторами: $\theta^i = \{\theta_1^i, \theta_2^i, \dots, \theta_k^i\}$, где ω – вектор весов, связанных с индивидуальными гауссовыми функциями; σ^i – вектор среднеквадратичных отклонений, который вычисляется следующим образом:

$$\sigma_l^i = \xi \sum_{j=1}^k \frac{|\theta_j^i - \theta_l^i|}{k-1}.$$

Параметр $\xi > 0$ вносит эффект подобный норме испарения феромона в дискретном алгоритме муравьиной колонии. В непрерывном алгоритме вводится архив решений, представленный таблицей из k строк. Каждая строка состоит из трех частей: 1) найденное муравьем решение $\theta_l = \{\theta_l^1, \theta_l^2, \dots, \theta_l^N\}$; 2) ошибка, вычисленная нечеткой системой; 3) вес решения ω_l . Решения упорядочены в архиве по возрастанию ошибки. Вес ω_l решения θ_l вычисляется согласно формуле:

$$\omega_l = \frac{1}{gk\sqrt{2\pi}} e^{-\left(\frac{l-1}{\sqrt{2}gk}\right)^2},$$

где q – задаваемый параметр алгоритма.

При добавлении нового решения в архив худшее из них удаляется. Этот процесс аналогичен процессу испарения феромона в классическом алгоритме.

Прямой алгоритм муравьиной колонии [16]. Алгоритм предложен М. Kong и Р. Tian [17], муравей в данном алгоритме отвечает за вычисление значений закрепленного за ним параметра. Каждый i -й муравей создает свое решение, генерируя нормально распределенное действительное число $N(\mu_i, \sigma_i)$. В алгоритме используются два вида феромонов: первый связан с центрами нормальных распределений $\mu = \|\mu_1, \dots, \mu_M\|$, второй – с разбросом $\sigma = \|\sigma_1, \dots, \sigma_M\|$. Количество феромона определяет значения параметров μ и σ . Для каждого параметра θ_j задан интервал изменения $[a_j, b_j]$, где b_j и a_j – верхняя и нижняя граница параметра θ_j .

В качестве начальных значений для параметров μ используются заданные

случайным или иным способом значения параметров θ . Начальные значения параметров σ вычисляются следующим образом:

$$\sigma_i = \frac{b_i - a_i}{2}.$$

После того как муравьи нашли решения, определяется испарение феромона. Для текущей t -й итерации испарение определяется следующим образом:

$$\mu(t) = (1 - \rho) \mu(t-1), \quad \sigma(t) = (1 - \rho) \sigma(t-1),$$

где ρ – эмпирический коэффициент, заданный на интервале $[0, 1]$.

Далее происходит нанесение феромона:

$$\mu(t) = \mu(t) + \rho\theta(t), \quad \sigma(t) = \sigma(t) + \rho|\theta(t) - \mu(t)|,$$

где $\theta(t)$ – решение, найденное муравьиной колонией на текущей итерации, оно совпадает с глобальным лучшим решением.

Для преодоления локальных минимумов в алгоритме используется обновление параметров σ . С этой целью введен параметр конвергенции, вычисляемый по следующей формуле:

$$cf = \frac{\sum_{j=1}^N \frac{2\sigma_j}{b_j - a_j}}{N}.$$

Когда алгоритм приближается к локальному минимуму, коэффициент конвергенции cf приближается к 0. Как только коэффициент конвергенции становится меньше критического значения cf_r , вектор σ возвращается в начальное состояние.

Эксперимент и обсуждение результатов

Исследование проводилось при решении задач аппроксимации идеальных данных и данных, описывающих реальные процессы. В качестве идеальных тестовых данных были выбраны две функции с двумя и тремя переменными:

- 1) $f(x_1, x_2) = \sin(2x_1/\pi) * \sin(2x_2/\pi)$, $-5 < x_1, x_2 < 5$;
- 2) $f(x_1, x_2, x_3) = 1 + x_1^{0.5} x_2^{-1} + x_3^{-1.5}$, $1 < x_1, x_2, x_3 < 5$.

На основе тестовой функции формировалась таблица наблюдений, по которой строилась нечеткая система, аппроксимирующая данную функцию. Исследование было проведено и на реальных данных, представленных в репозитории KEEL (Knowledge Extraction Evolutionary Learning, <http://www.keel.es>). Все входные и выходные переменные – вещественные числа. Каждая выборка разделена на пять наборов, из которых строится обучающая и тестовая выборки, содержащие 80% и 20% данных соответственно. Разделение проводилось таким образом, чтобы каждый набор попал во все тестовые выборки ровно один раз. Данные *diabetes* характеризуют проблему прогнозирования развития сахарного диабета у инсулинозависимых детей; количество наблюдений – 43; количество входных переменных – 2. Данные *ele-2* характеризуют проблему оценки стоимости обслуживания городских электрических сетей; количество наблюдений – 1066; количество входных переменных – 4.

Настройка параметров antecedентов правил проводилась гибридными методами. В состав гибридного алгоритма входили алгоритмы муравьиной колонии и дифференциальной эволюции, которые запускались последовательно один за другим. Оптимизация параметров консеквентов проводилась методом наименьших квадратов [19, 20].

На рис. 1 показаны усредненные значения информационных критериев для различных структур моделей при аппроксимации набора данных $f(x_1, x_2)$. Минимальное значение для всех критериев достигается при сложности равной 48, что соответствует модели с базой размером в 36 правил, в которой каждая входная переменная представлена шестью треугольными термами; при коэффициенте $c = 1$: $AIC = -11,901$, $BIC = -9,674$, $HQC = -11,376$; при коэффициенте $c = 3$: $AIC = -11,101$, $BIC = -6,678$, $HQC = -10,058$. Обоснованность выбора модели с такими параметрами подтверждается и минимальным значением ошибки MSE на тестовой выборке, которая чуть больше ошибки на обучающей выборке, что свидетельствует о высокой обобщающей способности и отсутствии переобучения.

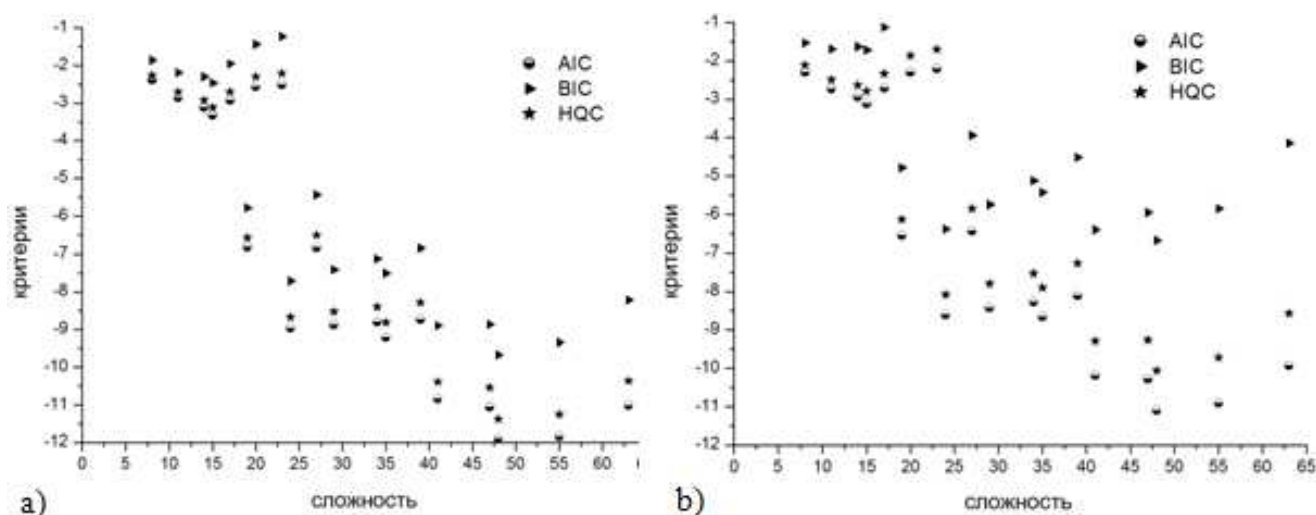


Рис. 1. Значения информационных критериев при аппроксимации набора данных $f(x_1, x_2)$: а) $c = 1$; б) $c = 3$.

На рис. 2 показаны усредненные значения информационных критериев при различных сгенерированных структурах модели для аппроксимации набора данных $f(x_1, x_2, x_3)$. Минимальное значение критериев при коэффициенте $c = 1$ достигается для сложности равной 59, что соответствует модели с базой размером в 48 правил, в которой одна входная переменная представлена тремя треугольными термами, а две другие – четырьмя термами; здесь $AIC = -7,889$, $BIC = -7,409$, $HQC = -7,705$. Если увеличить размер штрафа за правила, взяв $c = 9$, то выбор модели будет другим, не таким однозначным. Минимальные значения при $c = 9$ для критериев $AIC = -6,961$ и $HQC = -6,351$ достигаются на сложности равной 36, что соответствует модели с базой размером в 27 правил, в которой каждая входная переменная представлена тремя треугольными термами. Критерий BIC , который сильнее остальных двух критериев штрафует за лишние правила, принимает минимальное значение (равное $-5,470$) на сложности равной 14, что соответствует модели с базой размером всего в 8 правил, в которой каждая входная переменная представлена только двумя треугольными термами. В выбранных моделях ошиб-

ки MSE на тестовой выборке того же порядка, что и ошибки на обучающей выборке, этот факт свидетельствует об отсутствии переобучения.

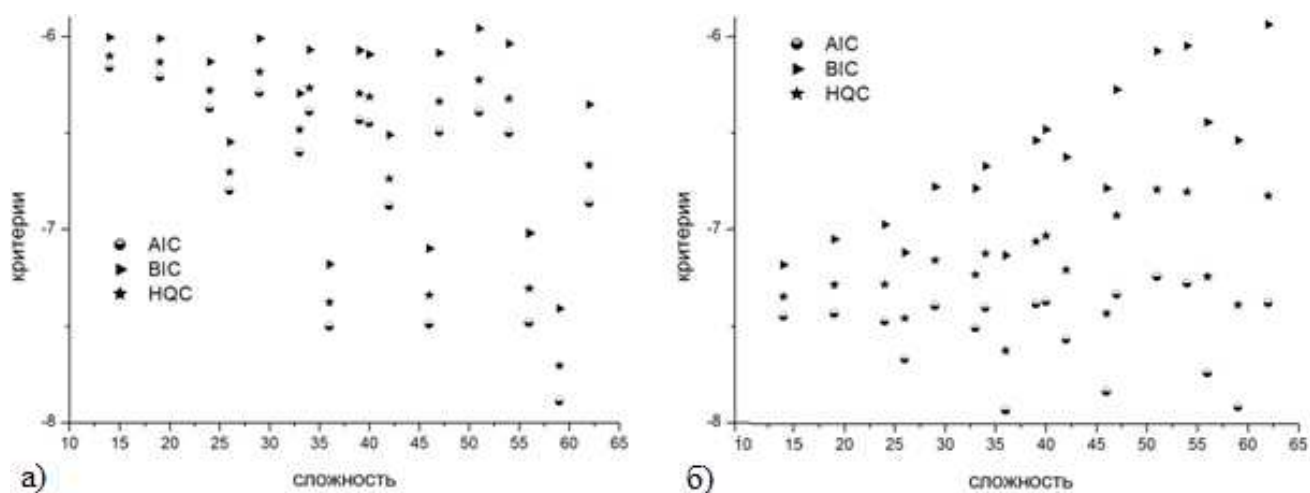


Рис.2. Значения информационных критериев при аппроксимации набора данных $f(x_1, x_2, x_3)$: а) $c = 1$; б) $c = 9$.

На рис. 3а показаны усредненные значения информационных критериев при различных сгенерированных структурах модели для аппроксимации набора данных *diabetes*. Менее всех штрафующий за лишние параметры модели критерий *AIC* указывает на модель сложности 15, в которой база правил имеет размер равный 9 и две входные переменные представлены тремя треугольными термами каждая. Критерии *BIC* и *HQC* предписывают выбрать модель минимальной сложности с базой размером всего в 4 правила, в которой каждая входная переменная представлена двумя термами.

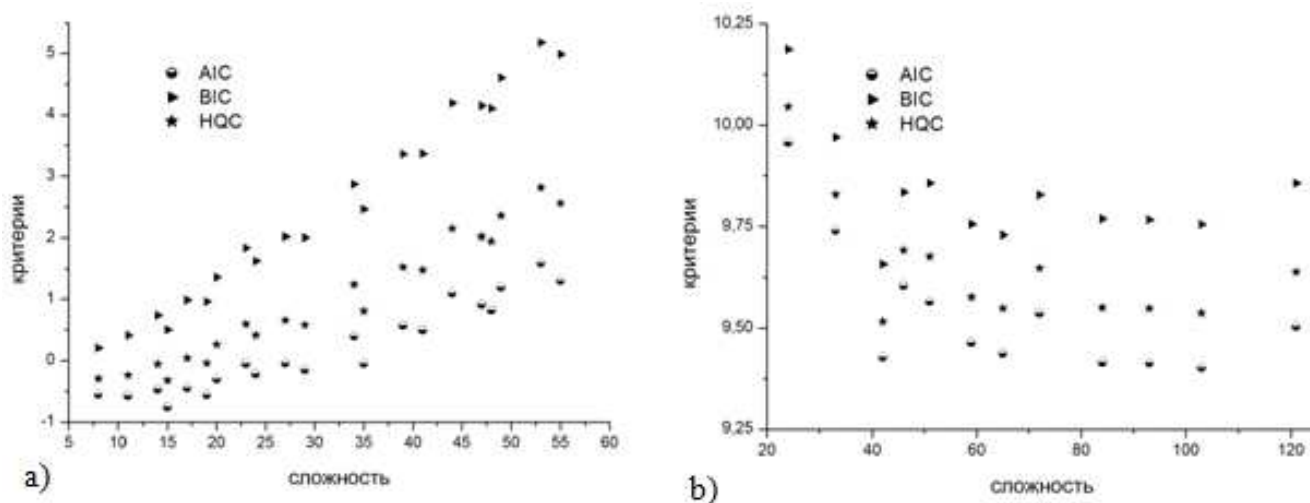


Рис.3. Зависимость «критерии – сложность» при аппроксимации реальных данных: а) *diabetes*; б) *ele-2*.

На рис. 3б показаны усредненные значения информационных критериев при различных сгенерированных структурах модели для аппроксимации набора данных *ele-2*. Критерий *AIC* предписывает выбрать модель сложности 103, в которой 90 правил, две входные переменные представлены тремя треугольными термами каждая, одна пятью и одна двумя термами. Критерии *BIC* и *HQC* здесь указывают на модель сложности 42, в которой база правил имеет размер равный

32, три входные переменные представлены двумя треугольными термами, а одна переменная – четырьмя.

Эксперимент показал, что не всегда все критерии указывают на единственную модель, в этом случае лицо принимающее решение должно выбрать из представленных моделей ту, которая, по его мнению, соответствует решаемой задаче.

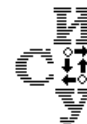
Заключение

В процессе построения нечеткой модели присутствуют два противоречивых требования: модель должна как можно точнее воспроизводить исследуемую систему, но должна быть простой и понятной для человека. Несложную модель не только легко понять, но также маловероятно, что эта модель будет переобучена. Однако невозможно построение модели одновременно с высокой степенью точности и понятности. В работе рассмотрен подход к нахождению компромисса между сложностью и точностью нечеткой модели.

Генерация структуры модели выполнена алгоритмом равномерного разбиения и перебора. Идентификация параметров antecedentов выполнена с помощью алгоритмов муравьиной колонии, параметры консеквентов правил настраиваются методом наименьших квадратов. Результаты эксперимента с идеальными и реальными данными показали пригодность подхода к поиску компромисса между сложностью и точностью модели с использованием информационных критериев.

ЛИТЕРАТУРА

1. *Gonzalez J., Rojas I., Pomares H., etc.* Improving the accuracy while preserving the interpretability of fuzzy function approximators by means of multi-objective evolutionary algorithms // International Journal of Approximate Reasoning. – 2007. – Vol. 44. – P.2-44.
2. *Guillaume S., Charnomordic B.* Learning interpretable fuzzy inference systems with FisPro // Information Sciences. – 2011. – Vol. 181. – P.4409-4427.
3. *Fazzolari M., Alcalá R., Nojima Y., etc.* A Review of the Application of Multiobjective Evolutionary Fuzzy Systems: Current Status and Further Directions // IEEE Trans. Fuzzy Systems. – 2013. – Vol. 21. – P.45-65.
4. *Ходашинский И.А., Горбунов И.В.* Алгоритмы поиска компромисса между точностью и сложностью при построении нечетких аппроксиматоров // Автометрия. – 2013. – Т. 49, № 6. – С.51-61.
5. *Hodashinsky I.A., Gorbunov I.V.* Algorithms of the Tradeoff between Accuracy and Complexity in the Design of Fuzzy Approximators // Optoelectronics, Instrumentation and Data Processing. – 2013. – Vol. 49, No. 6. – P.569-577.
6. *Ходашинский И.А., Горбунов И.В., Синьков Д.С.* Алгоритмы генерации структур двухкритериальных Парето-оптимальных нечетких аппроксиматоров // Доклады ТУСУР. – 2013. – №1(27). – С.135-142.
7. *Ходашинский И.А., Синьков Д.С.* Идентификация параметров нечетких систем на основе адаптивного алгоритма роящихся частиц // Информационные технологии. – 2011. – №8. – С. 2–5.
8. *Yen J., Wang L.* Application of Statistical Information Criteria for Optimal Fuzzy Model Construction // IEEE Trans. Fuzzy Systems. – 1998. – Vol. 6, N. 3. – P.362-372.
9. *Akaike H. A.* New Look at the Statistical Model Identification // IEEE Transactions on Automatic Control, AC-19. – 1974. – P.716-723.
10. *Schwarz G.* Estimating the Dimension of a Model // The Annals of Statistics. – 1978. – Vol. 6. – P.461-464.



11. *Hannan E.J., Quinn B.G.* The Determination of the Order of an Autoregression // Journal of the Royal Statistical Society, B. – 1979. – Vol. 41. – P.190-195.
12. *Storn R., Price K.* Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces // Journal of Global Optimization. – 1997. – Vol. 11. – P.341–359.
13. *Ходашинский И.А., Дудин П.А.* Идентификация нечетких систем на основе метода дифференциальной эволюции // Доклады ТУСУР. – 2011. – №1(23). – С.178-183.
14. *Ходашинский И.А., Дудин П.А.* Идентификация нечетких систем на основе непрерывного алгоритма муравьиной колонии // Автометрия. – 2012. – Т. 48, № 1. – С.63-71.
15. *Khodashinskii I. A., Dudin A.* Identification of fuzzy systems using a continuous ant colony algorithm // Optoelectronics, Instrumentation and Data Processing. – 2012. – Т. 48, N. 1. – P.54–61.
16. *Socha K., Dorigo M.* Ant colony optimization for continuous domains // Europ. Journ. Operational Research. – 2008. – Vol. 185. – P.1155-1173.
17. *Ходашинский И.А., Дудин П.А.* Идентификация нечетких систем на основе прямого алгоритма муравьиной колонии // Искусственный интеллект и принятие решений. – 2011. – №3. – С.26-33.
18. *Kong M., Tian P.* Application of ACO in Continuous Domain // ICNC 2006, Part II. LNCS 4222. – Berlin, Springer-Verlag, 2006. – P.126-135.
19. *Ходашинский И.А.* Идентификация нечетких систем на базе алгоритма имитации отжига и методов, основанных на производных // Информационные технологии. – 2012. – №3. – С.14-20.
20. *Ходашинский И.А.* Методы мягкого оценивания величин. – Томск: ТГСУиР, 2007.

Статья представлена к публикации членом редколлегии А.А. Шелупановым.

E-mail:

Ходашинский Илья Александрович – hodashn@rambler.ru.

УДК 621.372.542

© 2014 г. **А.Г. Шоберг**, канд. техн. наук
(Тихоокеанский государственный университет, Хабаровск)

МОДИФИЦИРОВАННОЕ ДИСКРЕТНОЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЕ ДВУХМЕРНОГО СИГНАЛА НА ОСНОВЕ БАЗИСА ХААРА

Рассматриваются подходы к организации работы с прикладным программным обеспечением в распределенных вычислительных системах, с применением проблемно-ориентированных интерфейсов. Приведено описание архитектуры разработанной программной платформы, в том числе сервисов контроля состояния инфраструктуры, и пример адаптации пакета *fhi98md* для работы в Грид.

Ключевые слова: распределенные вычислительные системы, проблемно-ориентированные интерфейсы, мониторинг сети.

Введение

Снижение количества вычислений при исследовании объектов и систем может быть достигнуто за счет инвариантности параметров по отношению к не-