УДК 519.25+517.11

ОБ ОДНОМ ПОДХОДЕ К ФОРМИРОВАНИЮ БАЗЫ ЗНАНИЙ ДЛЯ СЕГМЕНТАЦИИ ВРЕМЕННЫХ РЯДОВ

Т. М. Леденева, М. А. Сергиенко

Воронежский государственный университет

Поступила в редакцию 15.09.2017 г.

Аннотация. Предполагается, что поведение сложной системы описывается совокупностью временных рядов. С учетом их особенностей разработан комплекс подходов, базирующихся на статистическом анализе данных, алгоритмах кластеризации, методике нечеткого моделирования, который позволяет сформировать базу знаний для нечеткой системы прогнозирования поведения системы в форме продукционных правил с различными типами заключений.
Ключевые слова: нечеткая система, продукционное правило, кластер, форма кластера.
Annotation. It is assumed that the behavior of a complex system is described by a set of time series. In view of their features, a set of approaches has been developed. It is based on statistical analysis of data, algorithms for clustering of data, fuzzy modeling techniques and allows to create a knowledge base for a fuzzy system for forecasting the behavior of the system in the form of production rules with different types of conclusions.
Keywords: fuzzy system, if-then rule, cluster, cluster form.

ВВЕДЕНИЕ

При разработке моделей и методов анализа данных, результаты которых должны интегрироваться с системами поддержки принятия решений на основе знаний, центральное место занимают проблемы разработки адекватных методов моделирования компонент экспертных знаний, определенных на промежутках временных рядов.

Пусть в моменты времени $t = 1,...,T$ наблюдаются n параметров некоторой сложной системы, так что в каждый момент времени t состояние системы характеризуется вектором $x^t = \left(x_1^t,...,x_n^t\right)$. С другой стороны, каждому параметру $P_i\left(i = \overline{1,n}\right)$ можно поставить в соответствие временной ряд $X_{i,T} = \left\{x_i^1,...,x_i^T\right\}$. Таким образом, в промежутке $\left[1,...,T\right]$ поведение системы описывается совокупностью из n временных рядов.

Заметим, что если получить разбиение множества состояний $\left\{x^t\right\}_{t=\overline{1,T}}$ на m кластеров, когда в один кластер объединяются близкие в некотором смысле векторы состояний, то можно выделить m типовых состояний системы $S_j\left(j = \overline{1,m}\right)$, каждое из которых целесообразно описать в терминах ее параметров $P_i\left(i = \overline{1,n}\right)$.

Постановка задачи моделирования временного ряда в общем виде заключается в выявлении зависимости соответствующего ему показателя от времени на основе ретроспективных данных, а также определении оценки значения данного показателя в следующие моменты времени с учетом заданного горизонта прогнозирования, при этом предполагается, что выявленная зависимость сохранится на ограниченном отрезке времени в будущем. В настоящее время можно выделить несколько основных подходов к моделированию временных рядов: статистический [1, 2], нейросетевой [3, 4], нечеткий [5, 6]. Статистический подход, ставший классическим, основывается на восстановлении зависимости в форме статистической модели, которая включает систематическую и случайную составляющие, при этом систематическая составляющая в общем случае

является линейной комбинацией трендовой, периодической и сезонной компонент, а также (при необходимости) авторегрессионной компоненты. Моделирование временных рядов в рамках нейросетевого подхода сводится к задаче наилучшей аппроксимации нелинейной функции от многих переменных, параметрической моделью которой служит нейронная сеть. Возможность обучения – одно из главных преимуществ нейронных сетей, причем в процессе обучения нейронная сеть способна выявлять сложные нелинейные зависимости и выполнять обобщение. В основе нечеткого моделирования временных рядов лежит Fuzzy Approximation Theorem, согласно которой с помощью условных высказываний *если-то* с последующей их формализацией на основе лингвистических переменных, можно сколь угодно точно описать произвольную взаимосвязь «входы-выход».

Цель статьи заключается в представлении подхода для формирования базы знаний нечеткой системы прогнозирования состояния сложной системы на основе временных рядов, описывающих ее поведение на заданном временном промежутке. Центральной проблемой подхода является повышение качества аппроксимации, прежде всего, процедур кластеризации, которые являются базовыми для предложенного подхода. Решение данной проблемы осуществляется за счет разработки математической модели определения оптимальной формы кластера, суть которой заключается в выявлении наиболее подходящей формы кластера в виде кривой второго порядка на основе аппроксимации соответствующей матрицы квадратичной формы, что позволяет получить компактные кластеры. Каждому кластеру ставится в соответствие продукционное правило, что позволяет сгенерировать минимально необходимое количество правил, при этом за счет оптимизации формы кластеров улучшается точность аппроксимации, а, следовательно, качество обработки данных. Минимизация количества правил в базе знаний приводит к сокращению времени, необходимого для прогноза, и, в конечном счете, к ускорению обработки данных. Кроме того, база знаний, содержащая продукционные правила, учитывающие оптимальную форму кластеров, повышает объяснительные способности механизма нечеткого логического вывода, что делает прогноз в большей степени объяснимым и прозрачным.

1. ИСПОЛЬЗУЕМЫЕ МЕТОДЫ И ПОДХОДЫ

1.1. Основные сведения о квадратичных формах

Среди множества алгоритмов кластеризации перспективными в рамках рассматриваемой проблемы являются алгоритмы, формирующие кластеры определенной формы. Например, метод Густавсона-Кесселя [7] позволяет выделять кластеры в форме эллипсоидов, а нечеткий алгоритм кластеризации С-средних [8] продуцируют кластеры сферической формы. Существенным недостатком данных алгоритмов является то, что зачастую данным навязывается, возможно, не присущая им структура, а это может привести к принципиально неверным результатам. Поэтому актуальной проблемой является разработка алгоритмов, позволяющих извлекать из данных кластеры различной геометрической формы.

Основная идея подхода заключается в описании кластеров с помощью кривых, соответствующих различным матрицам квадратичной формы.

Введем основные определения [9].

Квадратичной формой $f\left(y_1,...,y_K\right)$ от переменных $y_1,...,y_K$ называется многочлен вида

$$f\left(y_1,...,y_K\right)=y^T Ay=\sum_{i=1}^{K}\sum_{j=1}^{K}a_{ij}y_i y_j= \\ =\sum_{i=1}^{K}a_{ii}^2 y_i^{\,2}+2\sum_{1\le i<j\le K}a_{ij}y_i y_j, \quad (1)$$

где y^T – вектор-строка, y – вектор-столбец переменных и не все коэффициенты a_{ij} нулевые.

Если для всех пар индексов (i,j) имеет место $a_{ij}=a_{ji}$, то квадратичная форма называется симметрической.

Для каждой действительной симметрической квадратичной формы существует действительная ортогональная матрица L, приводящая матрицу A к диагональному виду. Получающееся в результате преобразование к главным осям позволяет привести квадратичную форму к сумме квадратов

$$f(y_1,...,y_K) = y^T Ay = \xi^T L\xi = \sum_{i=1}^{n} \lambda_i \xi_i^2, \quad (2)$$

где $\{\lambda_1,...,\lambda_n\}$ – спектр матрицы A.

Преобразование $Y_i = \xi_i \Big/ \sqrt{|\lambda_i|}$ приводит (2) к нормальному виду

$$f(y_1,...,y_K) = \sum_{i=1}^{K} \varepsilon_i Y_i^2, \quad (3)$$

где каждый коэффициент ε_i равен $+1, -1$ или 0, если соответствующее собственное значение λ_i соответственно положительно, отрицательно или равно нулю.

Таким образом, любая квадратичная форма представима в виде

$$\begin{aligned} f(y_1,...,y_K) &= f(Y_1,...,Y_{K'}) = \\ &= Y_1^2 + ... + Y_s^2 - Y_{s+1}^2 - ... - Y_{K'}^2, \end{aligned} \quad (4)$$

где $K' \le K$, причем данное представление единственно.

Гиперповерхностью второго порядка в аффинном пространстве называется множество решений уравнения

$$f(y_1,...,y_K) + 2\alpha(y_1,...,y_K) + \alpha_0 = 0, \quad (5)$$

где $f(y_1,...,y_K)$ – квадратичная форма, $\alpha(y_1,...,y_K)$ – линейная или нулевая форма, α_0 – действительное число.

С учетом (4) уравнение (5) в новых координатах запишется в виде

$$Y_1^2 + ... + Y_s^2 - Y_{s+1}^2 - ... - Y_{K'}^2 = \begin{cases} 1 \\ -1 \\ 0 \\ 2X_{K'+1} \end{cases},$$

где $s \ge K' - s$; s – количество положительных собственных значений; K' – количество ненулевых собственных значений.

На плоскости линия второго порядка задается уравнением вида

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0,$$

где $A^2 + B^2 + C^2 \ne 0$.

При подходящем выборе системы координат рассматриваемое уравнение можно привести к одному из видов, представленных в табл. 1.

Таблица 1

№	Уравнение	Комментарий
1	$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1$	**Эллипс**
2	$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = -1$	Уравнение определяет мнимую линию второго порядка.
3	$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 0$	Уравнение задает начало координат.
4	$\frac{X^2}{a^2} - \frac{Y^2}{b^2} = 1$	**Гипербола**
5	$\frac{X^2}{a^2} - \frac{Y^2}{b^2} = 0$	Уравнение определяет пару пересекающихся прямых.
6	$Y^2 = 2pX$	**Парабола**
7	$Y^2 - a^2 = 0$	Уравнение определяет пару параллельных прямых.
8	$Y^2 + a^2 = 0$	Уравнение определяет мнимую линию второго порядка
9	$Y^2 = 0$	Уравнение определяет пару совпадающих прямых.

Уравнения 1, 4 и 6 определяют невырожденные кривые второго порядка.

Идея подхода заключается в выявлении наиболее подходящей формы кластера в форме кривой второго порядка на основе аппроксимации соответствующей матрицы квадратичной формы. Для нахождения соответствующих матриц каждого кластера и каждой пары координат решается оптимизационная задача, максимизирующая качество разбиения.

1.2. О форме кластера

Известно, что форма кластеров зависит от нормы, используемой для расчета расстояния между элементами. В общем случае норма за-

дается с помощью симметричной положительно определенной матрицы B в виде [7]

$$\rho_B(x,y) = \|x-y\|_B^2 = (x-y)\cdot B\cdot (x-y)^T, \quad (4)$$

где x, y – n-мерные векторы, B – матрица размерности $n \times n$.

Для евклидовой нормы матрица B представляет собой единичную матрицу. Для диагональной нормы – матрицу, на главной диагонали которой стоят веса координат. Диагональная норма позволяет выделять кластеры в виде эллипсоидов, ориентированных вдоль координатных осей. Для нормы Махаланобиса $B = R^{-1}$, где R – ковариационная матрица. Норма Махаланобиса позволяет выделять кластеры в виде эллипсоидов, оси которых могут быть ориентированы в произвольных направлениях.

Эллипсоид – это геометрическое место x всех точек, которые удовлетворяют условию

$$\alpha^2 = (x-c)^T B(x-c) = (x-c)^T P\Lambda P^T (x-c),$$

где α – положительное действительное число; c – центр эллипсоида; Λ – диагональная матрица с собственными значениями $\lambda_1,...,\lambda_q$ матрицы B на диагонали; P – ортогональная матрица, чьи столбцы являются собственными векторами $e_1,...,e_q$ матрицы B. Собственные вектора являются единичными и ортогональными и определяют оси эллипсоида. Матрица P задает поворот системы координат собственных векторов для ориентации эллипсоида. Собственные значения определяют длину осей.

На плоскости матрица квадратичной формы представима в виде

$$B = PDP^T, \quad (5)$$

где $D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ – матрица, на диагонали которой стоят собственные значения матрицы B, $P = \begin{pmatrix} e_1^1 & e_1^2 \\ e_2^1 & e_2^2 \end{pmatrix}$ – матрица единичных собственных векторов матрицы B, которая задает угол поворота системы координат Θ, то есть $P = \begin{pmatrix} \cos\Theta & \sin\Theta \\ \sin\Theta & -\cos\Theta \end{pmatrix}$. В этой системе координат уравнение квадратичной формы является каноническим $f(Y_1, Y_2) = \lambda_1 Y_1^2 + \lambda_2 Y_2^2$, где Y_1, Y_2 – переменные в новой системе координат.

Предположим, что матрица B, задающая форму кластера, известна. Для того чтобы определить, какой кривой соответствует квадратичная форма, необходимо найти ее собственные значения λ_1 и λ_2, а затем на их основе следующим образом определить тип кривой [10]:

a) если $\lambda_1 > 0, \lambda_2 > 0$, то кривая представляет собой эллипс (в нашем случае радиус положительный, поэтому случай мнимого эллипса и пары мнимых прямых можно не рассматривать);

b) если $\lambda_1 \cdot \lambda_2 < 0$, то кривая представляет собой гиперболу (пару пересекающихся прямых по той же причине не рассматриваем);

c) если $\lambda_1 = 0$ или $\lambda_2 = 0$, то квадратичная форма задает пару параллельных прямых.

2. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

2.1. Модель для определения геометрической формы кластера

Пусть получено разбиение n векторов исходных данных на кластеры $\{S_i\}_{i=\overline{1,m}}$, c^i – центр кластера S_i. Описание нечетких кластеров представлено матрицей нечеткого разбиения $\mu = \{\mu_{ij}\}_{n\times m}$, в которой i-я строка содержит степени принадлежности объекта x^i к кластерам $S_1,...,S_m$. Для каждой пары индексов (i,j) $\mu_{ij} \in [0,1]$ и выполняются следующие условия:

$$\sum_{j=1}^{m} \mu_{ij} = 1 \text{ для } i = \overline{1,n},$$

$$0 < \sum_{i=1}^{n} \mu_{ij} \le m \text{ для } j = \overline{1,m}.$$

Первое условие означает, что объект x^i принадлежит хотя бы одному классу, а второе – в каждом классе находится хотя бы один объект.

В соответствии с формулой (4) близость объекта x^j к центру c^i кластера S^i определим следующим образом:

$$\rho_{B_i}(x^j, c^i) = \|x^j - c^i\|_B^2 = (x^j - c^i)\cdot B_i \cdot (x^j - c^i)^T. \quad (6)$$

Требуется определить геометрическую форму в наибольшей степени соответствую-

щую каждому из кластеров. Для этого необходимо для каждого кластера S_i найти матрицу квадратичной формы B_i. Заметим, что значение $\rho_{B_i}\left(x^j, c^i\right)$ может быть отрицательным (так как собственные значения могут быть комплексными).

С учетом того, что степень принадлежности объекта x^j кластеру S_i есть μ_{ij}, и кластер должен быть компактным, рассмотрим следующую задачу для нахождения матрицы B_i:

$$F\left(S_i, B_i\right) = \\ = \sum_{\left\{j:\, j=\overline{1,N} \wedge \mu_{ij} > 0.5\right\}} \mu_{ij} \cdot \left(\rho_{B_i}\left(x^j, c^i\right)\right)^2 \to \min. \quad (7)$$

Рассмотрим задачу (7) на плоскости. Согласно (5), $B_i = P_i D_i P_i^T$, тогда с учетом определения матриц D_i и P_i переменными являются Θ, λ_1, λ_2. Для решения задачи (7) можно использовать метод градиентного спуска [11]. Частные производные целевой функции по каждой переменной имеют вид

$$\frac{\partial F}{\partial \Theta} = \sum_{\left\{j:\, j=\overline{1,n} \wedge \mu_{ij} > 0.5\right\}} 2\mu_{ij}\rho_{B_i}\left(x^j, c^i\right)\frac{\partial \rho_{B_i}}{\partial \Theta},$$

$$\frac{\partial F}{\partial \lambda_1} = \sum_{\left\{j:\, j=\overline{1,n} \wedge \mu_{ij} > 0.5\right\}} 2\mu_{ij}\rho_{B_i}\left(x^j, c^i\right)\frac{\partial \rho_{B_i}}{\partial \lambda_1},$$

$$\frac{\partial F}{\partial \lambda_2} = \sum_{\left\{j:\, j=\overline{1,n} \wedge \mu_{ij} > 0.5\right\}} 2\mu_{ij}\rho_{B_i}\left(x^j, c^i\right)\frac{\partial \rho_{B_i}}{\partial \lambda_2}.$$

Частные производные функции ρ_{B_i} по каждой из переменных имеют следующий вид:

$$\frac{\partial \rho_{B_i}}{\partial \Theta} = 2\cos\Theta\sin\Theta\left(x_p^j - c_p^i\right)^2\left(\lambda_2 - \lambda_1\right) + \\ +2\left(x_p^j - c_p^i\right)\left(x_q^j - c_q^i\right)\left(\lambda_1 - \lambda_2\right)\left[\cos^2\Theta - \sin^2\Theta\right] + \\ +2\cos\Theta\sin\Theta\left(x_p^j - c_p^i\right)^2\left(\lambda_1 - \lambda_2\right),$$

$$\frac{\partial \rho_{B_i}}{\partial \lambda_1} = \left(x_p^j - c_p^i\right)^2\cos^2\Theta + \\ +2\left(x_p^j - c_p^i\right)\left(x_q^j - c_q^i\right)\cos\Theta\sin\Theta + \\ +\left(x_{jq} - v_{iq}\right)^2\sin^2\Theta,$$

$$\frac{\partial \rho_{B_i}}{\partial \lambda_2} = \left(x_p^j - c_p^i\right)^2\sin^2\Theta - \\ -2\left(x_p^j - c_p^i\right)\left(x_q^j - c_q^i\right)\cos\Theta\sin\Theta + \\ +\left(x_p^j - c_p^i\right)^2\cos^2\Theta.$$

Подставляя значения частных производных в формулы для поиска решения по методу градиентного спуска [11], получим оптимальные значения переменных Θ^*, λ_1^*, λ_2^*, которые позволяют определить вид кривой второго порядка для данного кластера и данной пары координат (p, q).

Заметим, что задача (7) может рассматриваться для случая, когда $\rho_{B_i}\left(x^j, c^i\right)$ интерпретируется как расстояние [10].

Для распознавания новых объектов для каждого кластера S_i необходимо определить такое значение параметра r_i, для которого $\rho_{B_i}\left(x^j, c^i\right) \le r_i^2$ при $\mu_{ij} > 0.5\left(j = \overline{1, N}\right)$. Для эллипсов в качестве такого значения в методе Густавсона-Кесселя используется $r_i = \sqrt{\det\left(B_i\right)}$ [7]. В общем случае значение r_i можно подбирать экспериментально.

2.2. Формирование правил на основе кластеров эллипсоидальной формы

Пусть на плоскости кластер имеет форму эллипса, тогда ему соответствует «заплатка», которая как бы покрывает график функции, характеризующей зависимость выходной переменной от входной. На рис. 1 показаны эллипсы и их проекции на оси для функции с одним входом и одним выходом. На рис. 1 проекция эллипса на каждую из осей пространства состояния «вход-выход» определяет нечеткое треугольное число.

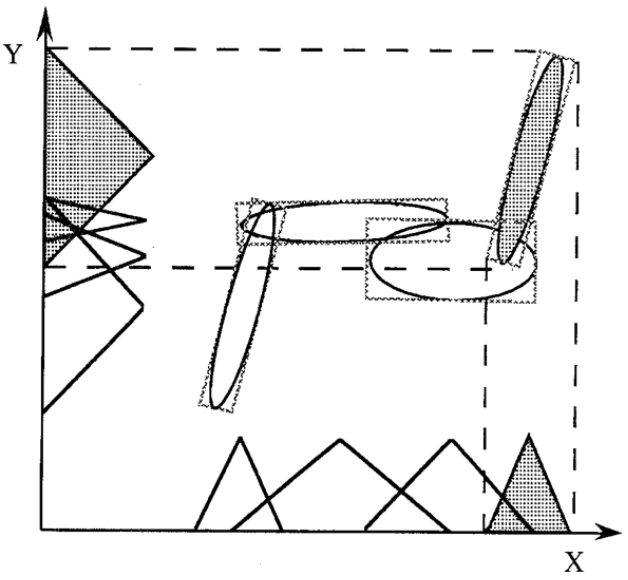

Рис. 1

Рассмотрим кластер S_k с центром c^k. Собственные вектора и собственные значения матрицы B_k определяют эллипсоид в q-мерном пространстве состояний «вход-выход» ($q = n + m$). Действуя согласно [12], впишем эллипсоид в гиперпараллелепипед, а затем спроецируем его на оси пространства состояний. Ориентация собственных векторов определяет размер проекции. k-й гиперпараллелепипед имеет 2^q вершин $(\pm\alpha_k / \sqrt{\lambda_{k_1}}, ..., \pm\alpha_k / \sqrt{\lambda_{k_q}})$ в повернутой плоскости координат. Единичное собственное значение определяет направление косинуса для каждой оси эллипсоида. Направление косинуса $\cos\gamma_{k_{ij}}$ – это угол между j-м собственным вектором и i-й осью k-го эллипсоида. Проекция k-го гиперпараллелепипеда на i-ю ось отцентрирована относительно c_{k_i} по i-ой оси и имеет длину

$$\rho_{k_q} = 2\alpha_k \sum_{j=1}^{q} \frac{\left|\cos\gamma_{k_{ij}}\right|}{\sqrt{\lambda_{k_j}}}.$$

На плоскости проекции прямоугольников имеют вид

$$\rho_{k_1} = 2\alpha_k \left(\frac{\left|\cos\theta\right|}{\sqrt{\lambda_{k_1}}} + \frac{\left|\sin\theta\right|}{\sqrt{\lambda_{k_2}}} \right),$$

$$\rho_{k_2} = 2\alpha_k \left(\frac{\left|\sin\theta\right|}{\sqrt{\lambda_{k_1}}} + \frac{\left|\cos\theta\right|}{\sqrt{\lambda_{k_2}}} \right).$$

Если выборка данных нечастая или шумная, то ковариация кластеров (заплаток) велика, в результате чего получаются большие по размеру эллипсоиды. Если данные плотные, то ковариация кластеров мала, в результате чего, получаются эллипсоиды малых размеров и более точные нечеткие правила. Большее количество заплаток лучше покрывает изгибы функции. Размер заплаток уменьшается с ростом их количества. Это порождает более точные и менее нечеткие правила. Проекции эллипсоидов должны покрывать область определения функции так, чтобы функция была определена для всех входов. Пересечение нечетких заплаток сглаживает аппроксимацию функций. Выбор α определяет, как сильно эллипсоиды будут пересекаться.

Заметим, что если отношение большей полуоси к меньшей достаточно велико, то целесообразно для аппроксимации использовать одну из осей эллипсоида. Тогда в заключении правила будет стоять линейная функция, а соответствующая модель нечеткой системы называется моделью Такаги-Сугено (или TS-моделью). Того же типа модель можно получить, если в рассматриваемом кластере восстановить функцию линейной регрессии. Необходимость в линейном заключении может возникнуть в связи с необходимостью установления тренда или приближенного (качественного) описания исследуемой зависимости.

2.3. Нечеткое моделирование временных рядов

В случае нечетких временных рядов в качестве модели авторегрессии используется реляционное уравнение вида [5]

$$y_{t+1}^j = y_t^i \circ R_{ij}\left(t+1,t\right),$$

где $y_{t+1}^j \in Y_{t+1}$, $y_t^j \in Y_t$, $i \in I$, $j \in J$; $\circ$ – операция композиции.

$R\left(t+1,t\right) = \bigcup\limits_{\{(i,j)\}} R_{ij}\left(t+1,t\right)$ – система нечетких отношений, которая формализует переход $Y_t \to Y_{t+1}$.

Предположим, что состояние системы в момент времени $\left(t+1\right)$ зависит от того, в каком состоянии она находилась в момент времени t. Пусть для формирования R используется некоторое обучающее множество Ω; $x^t = \left(x_1^t, ..., x_n^t\right)$ и $x^{t+1} = \left(x_1^{t+1}, ..., x_n^{t+1}\right)$ – векторы из Ω. Для удобства интерпретации значений параметров их приведем к безразмерной шкале $\left[0,1\right]$ с помощью подходящих функций нормирования.

Для моделирования перехода от x^t к x^{t+1} построим множество

$$\left\{\left(x_k^t, x_k^{t+1}\right)\right\}_{k=\overline{1,n}}$$

и будем рассматривать первые компоненты пар как значение входной переменной, а вторые компоненты – как значения выходной переменной.

Предположим, что с помощью подходящего алгоритма получено разбиение данного

множества на K кластеров. Заметим, что это разбиение также порождает разбиение показателей, поскольку в каждый кластер попадают точки $\left(x_k^t, x_k^{t+1}\right)$ из обучающего множества с определенными значениями индекса k. Таким образом, кластер отражает зависимость значений некоторой группы показателей, характеризующих состояние системы в момент времени $(t+1)$, от значений этих же показателей в предыдущий момент времени. Показатели, соответствующие данному кластеру, будем называть *существенными* для кластера.

Каждому кластеру можно поставить в соответствие продукционное правило, действующее для группы существенных показателей, причем заключение этого правила учитывает форму кластера. Количество правил определяется количеством кластеров. На основе приведенных выше рассуждений разработан следующий

Двухэтапный алгоритм синтеза базы знаний нечеткой системы, учитывающий форму кластеров

Вход: векторы, описывающий состояния системы в моменты времени t и $t+1$.

Выход: база знаний, состоящая из продукционных правил

Этап 1 (формирование кластеров оптимальной геометрической формы)

Замечание: в процессе работы алгоритма входная переменная всегда лингвистическая, в то время как выходная может быть числовой, лингвистической или ассоциированной с существенными параметрами рассматриваемого кластера.

1. Пусть в моменты времени t и $(t+1)$ состояния системы задаются векторами $x^t = \left(x_1^t, ..., x_n^t\right)$ и $x^{t+1} = \left(x_1^{t+1}, ..., x_n^{t+1}\right)$, причем $x_k^t, x_k^{t+1} \in [0,1]$ для всех $k = \overline{1,n}$. Сформировать обучающее множество в виде точек $\left\{\left(x_k^t, x_k^{t+1}\right)\right\}_{k=\overline{1,n}}$, при этом $x_k^t \left(k = \overline{1,n}\right)$ интерпретируются как числовые значения *входной* переменной X^t, а $x_k^{t+1} \left(k = \overline{1,n}\right)$ – как числовые значения *выходной* переменной X^{t+1}.

2. С помощью подходящего алгоритма нечеткой кластеризации сформировать матрицу нечеткого разбиения $\boldsymbol{\mu}$ и выделить кластеры $S_1, ..., S_K$.

3. Для повышения качества аппроксимации оптимизировать форму каждого кластера S_i на плоскости $X^t O X^{t+1}$ с помощью следующей процедуры:

3.1. Решая задачу (5), найти оптимальные значения Θ, λ_1, λ_2.

3.2. Проанализировать собственные значения и принять решение о форме кластера.

Этап 2 (формирование продукционных правил для каждого кластера)

4. Параметрам, входящим в кластер S_k, поставить в соответствие входную лингвистическую переменную $\tilde{X}_k^t$ и выходную лингвистическую переменную $\tilde{X}_k^{t+1}$ (при условии, что это необходимо для выбранного типа правила).

5. Каждому кластеру поставить в соответствие продукционное правило, при этом:

5.1. Если кластеру S_k соответствует эллипс, то определив проекции соответствующей матрицы квадратичной формы или проекции прямоугольника, описывающего эллипс, на оси OX^t и OX^{t+1}, сформировать функцию принадлежности $A_k(x)$ значения A_k лингвистической переменной $\tilde{X}_k^t$ и функцию принадлежности $B_k(x)$ значения B_k лингвистической переменной $\tilde{X}_k^{t+1}$, что позволяет получить правило вида

$$R_k : \textit{если } \tilde{X}_k^t \textit{ есть } A_k, \textit{ то } X_k^{t+1} = B_k.$$

5.2. Если кластер S_k представляет собой эллипс, «сильно вытянутый» вдоль одной из осей, то можно построить четкую или нечеткую функцию регрессии, что позволяет получить правило вида

$$R_k : \textit{если } \tilde{X}_k^t \textit{ есть } A_k, \textit{ то } \chi^{t+1} = \varphi_k\left(\chi^t\right),$$

где χ^t, χ^{t+1} – числовые переменные (соответственно входная и выходная), ассоциированные с существенными параметрами кластера S_k; $\varphi_k\left(\chi^t\right)$ – функция регрессии для соответствующих переменных.

5.3. Если кластер S_k описывается гиперболой, то соответствующее продукционное правило имеет вид

$$R_k : \textit{если } \tilde{X}_k^t \textit{ есть } A_k, \textit{ то } X_r^{t+1} = \frac{b}{a}\sqrt{\left(X_r^t\right)^2 - a^2},$$

где основной прямоугольник гиперболы ограничен прямыми $x = \pm a$, $y = \pm b$; $r \in I_k$.

5.4. Если кластер описывается прямой, то продукционное правило имеет вид

$$R_k : \text{если } \tilde{X}_k^t \text{ есть } A_k, \text{ то } X_k^{t+1} = \alpha X_k^t + \beta.$$

6. Проверить качество полученной базы правил. База правил и лингвистические шкалы для входных и/или выходных переменных составляют базу знаний.

Заметим, что к полученной базе правил при необходимости могут быть применены процедуры оптимизации [13]. Данный подход может быть обобщен на тот случай, когда состояние системы в момент времени $(t+1)$ зависит от ее состояния в предыдущие моменты времени $t,\ t-1,...,\ t-k$. В этом случае вместо пар необходимо рассматривать кортежи вида $\left\{\left(x_r^{t-k},...,x_r^t,x_r^{t+1}\right)\right\}_{r=\overline{1,n}}$.

ЗАКЛЮЧЕНИЕ

Предлагаемый подход актуален для прогнозирования поведения сложной системы в условиях неопределенности на основе анализа ретроспективных данных, полученных в результате наблюдения за определенный период времени. Его ключевой особенностью является наличие инструментов для формализации различных типов неопределенности – *физической*, обусловленной влиянием внешней среды или неточностью измерений, и *лингвистической*, возникающей вследствие использования естественного языка для оценки поведения и/или свойств сложной системы. В результате использования предложенной технологии обработки временных рядов формируется *сегментированный временной ряд*, представляющий собой совокупность сегментов, представленных кривыми определенного типа, словами, конкретным значением уровня временного ряда. В качестве обобщенной формы представления выступают продукционные (если-то) правила, действующие на конкретном временном промежутке. Сегментированный временной ряд может использоваться для решения задачи прогнозирования, а также следующих задач, относящихся к интеллектуальному анализу данных: кодирование временного ряда и уменьшения его размерности в случае big data; разработка методов представления и поиска в базах данных временных рядов, обеспечивающих быстрое выполнение запросов; выявление ассоциативных правил, связывающих группировки данных в соседних временных интервалах; дальнейший анализ временного ряда с целью выявления, например, аномалий, часто встречающихся группировок значений и т. д.

СПИСОК ЛИТЕРАТУРЫ

1. *Кендалл М., Стьюарт А.* Многомерный статистический анализ и временные ряды. – М. : Наука, 1976. – 736 с.

2. *Айвазян С. А.* Прикладная статистика и основы эконометрики / С. А. Айвазян, В. С. Мхитарян. – М. : ЮНИТИ, 1998.

3. *Каширина И. Л.* О методах формирования нейросетевых ансамблей в задачах прогнозирования временных рядов / И. Л. Каширина // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2009. – № 2.

4. *Леденева Т. М.* Обучение нейронных сетей методом Левенберга-Марквардта в условиях большого количества данных / Т. М. Леденева, С. С. Пархоменко // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2014. – № 2. – С. 8–106.

5. *Афанасьева Т. В.* Моделирование нечетких тенденций временных рядов / Т. В. Афанасьева. – Ульяновск : УлГТУ, 2013.

6. *Ярушкина Н. Г.* Метод нечеткого моделирования и анализа тенденций временных рядов / Н. Г. Ярушкина, Т. В. Афанасьева // Интеллектуальные системы управления. – М. : Машиностроение, 2010.

7. *Gustafson E.* Fuzzy clustering with a fuzzy covariance matrix/ E. Gustafson, W. Kessel // Proc. IEEE Conf. Decision Control, 1979.

8. *Anderberg M. R.* Cluster Analysis for Applications / M.R. Anderberg // Academic Press, New York. – 1973.

9. *Баскаков А. Г.* Лекции по алгебре / А. Г. Баскаков. – Воронеж : Изд-во ВГУ, 1999. – 284 с.

10. *Тарасова А. С.* Методы определения оптимальной геометрической формы в задачах

кластерного анализа / А. С. Тарасова // Информационные технологии. – 2007. – С. 14–21.

11. *Реклейтис Г.* Оптимизация в технике: в 2-х кн. Кн. 1 / Г. Реклейтис, А. Рейвиндран, К. Рэгсдел. – М. : Мир. 1986. – 350 с.

12. *Леденева Т. М.* Об одном подходе к идентификации нечетких систем / Т. М. Леденева, Д. С. Татаркин // Современные сложные системы управления: Сб. тр. Междунар. конф. (Воронеж, 2005г.). – Воронеж : ВГАСУ, 2005. – Т. 1. – С. 42–46.

13. *Леденева Т. М.* Оптимизация нечетких правил в задаче нечеткого прогнозирования / Т. М. Леденева, Д. С. Татаркин, А. С. Тарасова // Экономическое прогнозирование: модели и методы: матер. Междунар. науч.-прак. конф (Воронеж, 2007г.). – Воронеж : Воронеж. гос. ун-т, 2007. – С. 62–67.

Леденева Т. М. – д-р техн. наук, профессор, заведующий кафедрой вычислительной математики и прикладных информационных технологий, факультет прикладной математики, информатики и механики, Воронежский государственный университет.
E-mail: ledeneva-tm@yandex.ru

Сергиенко М. А. – канд. техн. наук, руководитель подразделения представительства АО «ОТ-ОЙЛ» в г. Воронеж.
E-mail: msergienko@ot-oil.com

Ledeneva T.M. – Doctor of Technical Science, Professor, Department of Computational Mathematics and Applied Information Technologies, Faculty of Applied Mathematics, Informatics and Mechanics, Voronezh State University.
E-mail: ledeneva-tm@yandex.ru

Sergienko M. A. – PHD in Technical Sciences, manager of reporting department «OT-OIL» Voronezh.
E-mail: msergienko@ot-oil.com