Минимизация числа состояний нечёткого автомата с помощью интервальных формальных понятий

Панкратьева В.В.[1]

В настоящей статье рассматривается связь между задачей минимизации числа состояний нечёткого автомата и проблемой поиска интервальных формальных понятий с максимальным объёмом. Метод кластеризации, основанный на поиске интервальных формальных понятий, позволяет объединить состояния нечёткого автомата в подмножества со сходными строками достоверностей перехода в другие состояния. При этом мера близости строк определяется заранее заданным параметром σ. В работе показано, что для определённого вида нечётких матриц перехода, начиная с некоторого момента поведение исходного нечёткого автомата, как и поведение минимизированного автомата, стабилизируется. Кроме того, доказано, что при минимизации автомата достоверность слова, распознаваемого нечётким автоматом, не уменьшается. Этот факт позволяет сравнить нечёткий язык, распознаваемый исходным автоматом, и язык, распознаваемый минимизированным.

Ключевые слова: нечёткие автоматы, интервальные формальные понятия, нечёткие языки.

1. Введение

Начиная со второй половины прошлого столетия активно исследуются различные математические модели, в которых некоторые параметры определены неоднозначно или не определены вовсе. Прекрасным инструментом для работы с такими моделями является созданная в 1965 году

[1] *Панкратьева Вера Викторовна* — аспирант каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: vera.pankratieva@gmail.com.

Pankratieva Vera Viktorovna — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

теория нечётких множеств [1]. Нечёткие методы применяются в различных прикладных разделах математики, в том числе в теории графов, теории формальных понятий, а также в теории автоматов и в других областях.

Конструкция нечётких автоматов впервые появилась в 1969 году в работе [2] и является обобщением понятия детерминированного автомата, в котором функция переходов в другое состояние не является однозначно определённой, а представляет собой некоторую оценку (степень нечёткости) из частично-упорядоченного множества. Чаще всего в качестве этого множества используется отрезок [0,1], в этом случае функция перехода представляет собой нечёткую матрицу, элементами которой являются числа из отрезка. Удобнее всего использовать не весь числовой отрезок, а значения из него с некоторым заданным постоянным шагом, например, шаг длиной 0.2 порождает шесть числовых значений: 0, 0.2, 0.4, 0.6, 0.8, 1. Для различных символов входа автомата такие нечёткие матрицы, вообще говоря, различны. В качестве операций умножения и сложения нечётких матриц используются стандартная операция min-max умножения и операция взятия максимума соответственно. Если входной алфавит представляет собой один символ, то его повторение отвечает возведению в степень нечёткой матрицы переходов.

Зачастую нечёткие матрицы имеют очень большие размеры, что вызывает неудобство при работе с ними, а также при интерпретации результатов работы модели. Одним из решений этой проблемы является минимизация числа состояний нечёткого автомата. Эта проблема может быть решена за счёт объединения в одно состояние нескольких «похожих» состояний нечёткого автомата. В нечёткой матрице переходов размера $n \times n$ состоянию с номером i соответствует строка степеней нечёткости, с которыми i-е состояние переходит во все остальные состояния автомата. Таким образом, если два или более состояний автомата имеют одинаковые или схожие строки степеней нечёткости в матрице, то такие состояния можно объединить в одно состояние, так как они ведут себя при функционировании автомата похожим образом. Данная задача полностью соответствует задаче нахождения интервальных формальных понятий.

Анализ формальных понятий — математическая теория, основы которой были заложены Б. Гантером и Р. Вилле в фундаментальной работе [3]. Анализ формальных понятий успешно применяется в различных системах кластеризации. При этом часто возникает необходимость работы с большим массивом данных, который представляет собой числовой контекст (таблицу). Существует несколько обобщений анализа формальных понятий именно для числовых и нечётких контекстов. Одно из обобщений — узорные структуры, введённые С. Кузнецовым и Б. Гантером [4], частным случаем которых являются формальные понятия с операцией интервального пересечения (слияния). Интервальные формальные понятия удобно использовать при работе с нечёткими контекстами для решения задач кластеризации, то есть разбиения множества строк на группы, определяемые похожими распределениями числовых данных.

Проблема поиска интервального формального понятия, содержащего наибольшее количество строк, NP-трудна [5], однако на практике для его нахождения может успешно применяться приближённый алгоритм [6].

Матрица переходов нечёткого автомата может интерпретироваться как нечёткий контекст, при этом задача уменьшения числа состояний автомата соответствует задаче кластеризации. В результате кластеризации получается новый нечёткий автомат, состояния которого соответствуют найденным кластерам.

Множество всех интервальных формальных понятий (найденных кластеров) образует полную решётку, что согласуется с результатами работы [7].

Одной из практически важных задач в теории нечётких автоматов является распознавание языков. Нечёткий автомат можно рассматривать как устройство, распознающее язык над входным алфавитом, при этом каждому слову ставится в соответствие степень его принадлежности языку.

Часто работа с данными требует применения таких моделей, которые бы позволяли выделить подмножество слов в языке, обладающих высокой (не ниже заранее заданного числа) степенью достоверности. В то же время степень достоверности распознаваемого автоматом слова может измениться при минимизации числа состояний исходного автомата. Поэтому естественным образом возникает задача о сравнении языков, распознаваемых до и после минимизации.

2. Основные понятия и определения

С детальным введением в теорию решёток и теорию порядков можно ознакомиться в работе [8], а основные определения, относящиеся к теории формальных понятий, содержатся в работе [3].

Определение 1. Бинарное отношение R на множестве M называется *отношением частичного порядка* (или просто *частичным порядком*), если оно удовлетворяет следующим условиям для всех элементов:

1) рефлексивность: xRx;
2) антисимметричность: если xRy и $x \neq y$, то не выполняется yRx;
3) транзитивность: xRy и $yRz \Rightarrow xRz$.

Для R мы часто используем символ $\leq$ (для R^{-1} символ $\geq$) и пишем $x < y$, если $x \leq y$ и $x \neq y$. Как обычно, мы читаем $x \leq y$ как «x меньше или равно y». *Частично упорядоченное множество* — это пара $(M, \leq)$, где M — множество, а $\leq$ — отношение частичного порядка на M.

Определение 2. *Верхняя (нижняя) полурешётка* — это частично упорядоченное множество $(M, \leq)$ такое, что для любых элементов $x, y \in M$ существует единственная точная верхняя (нижняя) грань.

Определение 3. *Полурешёточной операцией* на множестве M называется такая операция $\sqcap\colon M \times M \to M$, что для некоторого $e \in M$ и любых $x, y, z \in M$ выполнены следующие свойства:

- $x \sqcap x = x$ (идемпотентность);
- $x \sqcap y = y \sqcap x$ (коммутативность);
- $(x \sqcap y) \sqcap z = x \sqcap (y \sqcap z)$ (ассоциативность);
- $e \sqcap x = e$.

Определение 4. *Решёткой* называется упорядоченное множество $(L, \leq)$, которое является одновременно верхней и нижней полурешёткой.

Определение 5. Пусть $(P, \leq_P)$ и $(Q, \leq_Q)$ — частично упорядоченные множества. *Соответствием Галуа* между этими множествами называется пара отображений $\varphi\colon P \to Q$ и $\psi\colon Q \to P$ (каждое из которых является *оператором Галуа*) таких, что для любых $p_1, p_2 \in P$ и $q_1, q_2 \in Q$ выполнены следующие свойства:

- $p_1 \leq_P p_2 \;\Rightarrow\; \varphi(p_1) \geq_Q \varphi(p_2)$ (антиизотонность);
- $q_1 \leq_Q q_2 \;\Rightarrow\; \psi(q_1) \geq_P \psi(q_2)$ (антиизотонность);
- $p \leq_P \psi(\varphi(p))$ и $q \leq_Q \varphi(\psi(q))$ (изотонность).

Определение 6. Двойное применение оператора Галуа, а именно $\psi(\varphi(\cdot))$ и $\varphi(\psi(\cdot))$ является *оператором замыкания. Оператор замыкания* $\overline{(\cdot)}$ на M — отображение, относящее *замыкание* $\overline{X} \subseteq M$ каждому подмножеству $X \subseteq M$ и удовлетворяющее следующим условиям:

- $X \leq Y \ \Rightarrow\ \overline{X} \leq \overline{Y}$ (монотонность);
- $X \leq \overline{X}$ (экстенсивность);
- $\overline{\overline{X}} = \overline{X}$ (идемпотентность).

Сформулируем определение узорной структуры так, как оно дано в работе [4].

Определение 7. *Узорная структура* — это тройка $(G, (D, \sqcap), \delta)$, где G — множество объектов, $(D, \sqcap)$ — полная полурешётка всевозможных описаний, а $\delta \colon G \to D$ — функция, которая сопоставляет каждому объекту из множества G его описание из D.

Соответствие Галуа между подмножествами множества объектов и описаниями для узорной структуры $(G, (D, \sqcap), \delta)$ определяется следующим образом:

$$\begin{aligned} A^{\square} &:= \prod_{g \in A} \delta(g), && \text{где}\ \ A \subseteq G,\\ d^{\square} &:= \{g \in G \mid d \sqsubseteq \delta(g)\}, && \text{где}\ \ d \in D. \end{aligned}$$

Определение 8. *Узорное понятие узорной структуры* $(G, (D, \sqcap), \delta)$ — это пара (A, d), в которой $A \subseteq G$ — подмножество множества объектов, $d \in D$ — одно из описаний из полурешётки, такая, что $A^{\square} = d$ и $d^{\square} = A$; A называется *узорным объёмом* понятия, а d — его *узорным содержанием*.

Частным случаем узорного понятия является интервальное формальное понятие. В качестве D берутся строки числового контекста и рассматриваются как кортежи интервалов нулевой длины. Интервальное формальное понятие — это пара (A, d), где A — подмножество множества объектов, а d — кортеж интервалов, границы которых определяются наименьшим и наибольшим значениями соответствующей компоненты в описаниях всех объектов из A.

Интервальные формальные понятия удобно применять при работе с числовыми контекстами, когда необходимо разделять все данные на кластеры, содержащие объекты с похожими «распределениями» числовых данных в строке.

Если ввести параметр σ для ширины интервального понятия (разницы между наименьшим и наибольшим значениями каждой компоненты), то кластером можно будет назвать интервальное понятие с наибольшим объёмом, такое, что ширина по каждой компоненте не превосходит σ.

Далее приведём определения, относящиеся к теории нечётких автоматов, которые можно найти в работах [1], [2] и [9].

Определение 9. Пусть S — произвольное непустое множество элементов, $|S| = k$. *Нечётким подмножеством множества* S называется совокупность $\mu(S) = (\mu(s_1), \mu(s_2), \ldots, \mu(s_k))$, где $\mu(s_i) \in [0, 1]$, $1 \leq i \leq k$.

То есть для каждого элемента множества S указывается степень нечёткости или достоверности, с которой этот элемент принадлежит множеству S. На практике в качестве степеней нечёткости удобно использовать не произвольные точки отрезка [0, 1], а точки, полученные при делении отрезка на равные части, например, при делении на 10 *количество* α *степеней нечёткости* будет 11: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, а при делении на 5 значение α равно 6. Очевидно, что множество нечётких подмножеств множества S образует полную решетку.

Определение 10. *Нечёткая матрица* — это числовая матрица, элементы которой принадлежат отрезку [0,1].

Определение 11. *Нечётким языком* L *в алфавите* A называется произвольное нечёткое подмножество множества всех слов A^* в этом алфавите

$$L\colon A^* \to [0, 1].$$

Определение 12. *Нечёткий конечный инициальный автомат* — это четвёрка $F = (A, Q, \varphi, P_I)$, где $A = \{a_1, a_2, \ldots, a_m\}$ — конечный входной алфавит, $Q = \{q_1, q_2, \ldots, q_k\}$ — конечное множество состояний, $\varphi : Q \times A \times Q \to [0, 1]$ — функция переходов, P_I — вектор-строка начального состояния (нечёткое подмножество Q).

Запись $\varphi : Q \times A \times Q \to [0, 1]$ интерпретируется следующим образом: нечёткий автомат, находящийся в состоянии q_1, при действии входного символа a переходит в состояние q_2 с некоторой степенью нечёткости из отрезка [0, 1]. Эту степень нечёткости удобно обозначать как величину $\mu_a(q_1, q_2)$. Функция φ порождает для каждого входного символа a нечёткую матрицу $T(a)$ размера $k \times k$, $T(a) = [\mu_a(q_i, q_j)]$, где $1 \leq i \leq k$, $1 \leq j \leq k$, $a \in A$.

Определение 13. *Автономный автомат* — это автомат, входной алфавит которого состоит из одного символа, то есть $A = \{a\}$.

Для определения нечёткого языка, распознаваемого автоматом F, необходимо указать, каким образом вычисляется степень достоверности распознаваемого слова.

Пусть w — слово из A^* длины n. Слово w можно записать как последовательность входных символов $\omega_1\omega_2\ldots\omega_n$ из алфавита A. При подаче на вход одного из символов ω_i нечёткий автомат переходит из состояния q_j в состояние q_{j+1} с некоторой степенью достоверности $\mu_{\omega_i}(q_j, q_{j+1})$ из отрезка [0,1]. Тогда j-я «траектория» множества «траекторий» Ω автомата F при прочтении слова w выглядит следующим образом:

$$q_{j_0} \longrightarrow q_{j_1} \longrightarrow \ldots \longrightarrow q_{j_{n-1}} \longrightarrow q_{j_n},$$

где q_{j_0} — начальное состояние, а q_{j_n} — состояние автомата, в которое он попадает после прочтения слова w при прохождении по j-й «траектории». При этом последовательность изменений степеней достоверности перехода в следующее состояние выглядит как следующая цепочка:

$$\mu(q_{j_0})\to\mu_{\omega_1}(q_{j_0}, q_{j_1})\to\mu_{\omega_2}(q_{j_1}, q_{j_2})\to\ldots\to\mu_{\omega_{n-1}}(q_{j_{n-2}}, q_{j_{n-1}})\to\mu_{\omega_n}(q_{j_{n-1}}, q_{j_n}),$$

где $\mu(q_{j_0})$ — достоверность начального состояния j-й «траектории». Для j-й «траектории» можно ввести обозначение $M^j(w)=\min\limits_{0\le l\le n-1}\mu_{\omega_l}(q_{j_l}, q_{j_{l+1}})$, определяющее минимальное значение достоверности состояний нечёткого автомата при прочтении слова w.

Определение 14. *Степень достоверности распознаваемого автоматом F с помощью множества P_F слова w* определяется по формуле:

$$L(F)(w) = \max_j(\min(\mu(q_{j_0}), \mu(q_{j_n}), M^j(w))),$$

где $|w| = n$, q_{j_0} — начальное состояние для j-й «траектории», $\mu(q_{j_0})$ — достоверность начального состояния, q_{j_n} — состояние, в котором окажется нечёткий автомат при прохождении по j-й «траектории» после прочтения слова w, $\mu(q_{j_n})$ — соответствующий состоянию q_{j_n} элемент нечёткого подмножества $P_F{=}\mu(Q)$ множества состояний Q. Заметим, что начальное состояние P_I представляет собой нечёткую вектор-строку, а множество P_F можно интерпретировать как нечёткий вектор-столбец.

Другими словами, среди всех цепочек изменений состояний и достоверностей находится такая «траектория», что минимальная степень нечёткости (достоверности) перехода из одного состояния в другое максимальна среди всех возможных «траекторий». Очевидно, что $L(F)(w)$ принимает единственное значение, при этом «траекторий» из множества Ω, реализующих $L(F)(w)$, может быть много.

Таким образом, нечёткий автомат F с помощью множества финальных состояний P_F определяет степень достоверности каждого слова из A^*.

Определение 15. Множество распознаваемых автоматом F с помощью множества P_F слов с соответствующими им степенями достоверности называется *нечётким языком, распознаваемым автоматом F с помощью множества P_F*, и обозначается $L(F)$.

Иногда требуется найти подмножество слов из $L(F)$, у которых степень принадлежности не меньше какого-то наперёд заданного числа β. Такое подмножество нечёткого языка обозначается $L^{\beta}(F)$.

3. Минимизация числа состояний нечёткого автомата

Рассмотрим автономный нечёткий инициальный автомат $F=(A, Q, \varphi, P_I)$, где $A = \{a\}$, $Q = \{q_1, q_2, \ldots, q_k\}$ — конечное множество состояний, $\varphi : Q \times A \times Q \to [0, 1]$ — функция переходов, которая определена нечёткой матрицей перехода $T(A)$ размера $k \times k$, P_I — начальное распределение состояний (под словом "распределение" мы подразумеваем кортеж достоверностей элементов нечёткого подмножества, при этом, в отличие от теоретико-вероятностного значения этого термина, сумма достоверностей элементов нечёткого подмножества не обязательно равна 1).

На пересечении i-й строки и j-го столбца нечёткой матрицы перехода указана степень достоверности, с которой состояние q_i переходит в состояние q_j. То есть элементы i-й строки представляют собой степени достоверности, с которыми состояние q_i переходит во все состояния $\{q_1, q_2, \ldots, q_k\}$ при поступлении на вход символа a. Если в нечёткой матрице есть строки с таким же или в достаточной мере похожим распределением степеней достоверности, то можно все такие строки объединить в один кластер, то есть в одно новое состояние q'_i для нового нечёткого автомата $F' = (A, Q', \varphi', P'_I)$. Мера похожести двух строк m и n определяется выбранным параметром σ: строки похожи, если для любого столбца j, $1 \le j \le k$, выполняется неравенство $|(\mu_a(q_m, q_j)) - (\mu_a(q_n, q_j))| \le \sigma$.

Матрицу нечёткого перехода можно интерпретировать как нечёткий контекст, состояния которого $\{q_1, q_2, \ldots, q_k\}$ соответствуют множеству объектов G узорной структуры [4], а строки распределений, соответствующие состояниям — описаниями D для узорной структуры. Описание состояния q_i представляет собой кортеж длины k интервалов нулевой длины: $\{(\mu_a(q_i, q_1)), (\mu_a(q_i, q_2)), \ldots, (\mu_a(q_i, q_i)), \ldots, (\mu_a(q_i, q_k))\}$. Это соответствие определяет функцию $\delta : G \to D$, которая сопоставляет каждому объекту из множества G его описание из D.

Описанием для подмножества объектов J является кортеж длины k интервалов вида $\{\min\limits_{i \in J} \mu_a(q_i, q_l), \max\limits_{i \in J} \mu_a(q_i, q_l)\}$, где i — номера объектов (состояний), входящих в подмножество J, $1 \le l \le k$. Соответствие Галуа между подмножествами множества объектов (состояний) и множеством описаний (кортежей интервалов) для узорной структуры $(G, (D, \sqcap), \delta)$ определяется следующим образом:

$$J^{\Box} := \prod_{g \in J} \delta(g), \qquad \text{где } J \subseteq G,$$

$$d^{\Box} := \{g \in G \mid d \sqsubseteq \delta(g)\}, \qquad \text{где } d \in D.$$

Узорным понятием узорной структуры (в нашем частном случае интервальным формальным понятием) называется пара (J, d) такая, что $J^{\Box} = d$ и $d^{\Box} = J$. Эти формулы можно интерпретировать в следующем смысле: подмножеству множества объектов J (оно называется узорным объёмом) соответствует описание d (узорное содержание), причем подмножество J состоит из максимального числа объектов, разделяющих это описание, и это описание нельзя уточнить.

Таким образом, поиск похожих состояний нечёткого автомата сводится к поиску интервальных формальных понятий с максимальным узорным объёмом и с таким узорным содержанием, что длина интервала в каждой компоненте содержания не превышает значения σ. В качестве первого кластера выбирается интервальное формальное понятие, для которого выполнены эти условия; после выбора объекты удаляются из нечёткого контекста и процедура поиска интервального формального понятия с максимальным узорным объёмом и узорным содержанием ширины не более σ повторяется для меньшего нечёткого контекста.

После всех возможных итераций этой процедуры исходный нечёткий контекст будет представлять собой объединение непересекающихся кластеров, являющихся интервальными формальными понятиями. На языке нечётких автоматов это означает, что множество состояний Q будет поделено на непересекающиеся подмножества состояний нечёткого

автомата:
$$Q = \bigcup_i Q_i, \quad Q_i \cap Q_j = \varnothing, \quad i, j \in \{1, \ldots, k\}, i \neq j.$$
В каждом таком подмножестве Q_i содержатся состояния с похожими распределениями степеней достоверности и, таким образом, каждое подмножество Q_i станет состоянием q'_i для нового нечёткого автомата $F' = (A, Q', \varphi', P'_I)$. Значения новой нечёткой матрицы перехода $T'(A)$ определяются так, что на пересечении строки i и столбца j находится максимальная степень достоверности:
$$\mu_a(q'_i, q'_j) = \max_{q_m \in Q_i, q_n \in Q_j} \mu_a(q_m, q_n),$$
а значения нового вектора начального состояния P'_I определяются соотношением $\mu(q'_i) = \max\limits_{q_m \in Q_i} \mu(q_m)$.

В классической теории анализа формальных понятий существует множество алгоритмов для поиска замкнутых объектов, а также для построения решёток формальных понятий. Самые распространённые алгоритмы и их особенности описаны в статье [11]. Для поиска замкнутых объектов в многозначных и нечётких контекстах, а также в узорных структурах, существует множество модификаций этих известных алгоритмов.

Однако вычислительная сложность алгоритмов для поиска формальных понятий может быть экспоненциальной от числа объектов, это показано в работе [3]; кроме того, результатом работы [5] является факт, что задача поиска интервального формального понятия, содержащего максимальное количество объектов, NP-трудна. Поэтому на практике для поиска кластеров нечёткого контекста удобно использовать жадный алгоритм, приближенно решающий данную задачу за линейное по количеству объектов контекста и полиномиальное по размерности время. Приближенный алгоритм и оценки вычислительной сложности подробно описаны в работе [6]. Корректность работы алгоритма кластеризации была протестирована на данных [12], и результат работы приближенного алгоритма совпал с реальными данными.

В случае нечёткого автомата с произвольным конечным входным алфавитом $A = \{a_1, a_2, \ldots, a_{s-1}, a_s\}$ каждому входному символу a_i соответствует своя нечёткая матрица перехода $T(a_i)$. Для того, чтобы два состояния q_m и q_n нечёткого автомата попали в один кластер, необходимо, чтобы разность достоверностей в m-й и n-й строках не превышала значения σ: $|\mu_a(q_m, q_j) - \mu_a(q_n, q_j)| \leq \sigma$, $1 \leq j \leq k$, и это должно быть выполнено для каждой нечёткой матрицы $T(a_i)$. Как и в случае автономного автомата, нечёткие матрицы переходов можно рассматривать как нечёткие контексты. Если из всех матриц составить единую прямоугольную матрицу размера $k \times (k * s)$, (k строк и $k * s$ столбцов), то строки длиной $k * s$ можно рассматривать как описания узорной структуры, в которой состояния $\{q_1, q_2, \ldots, q_k\}$ соответствуют множеству объектов. В этой узорной структуре можно найти интервальные формальные понятия с максимальным узорным объёмом и шириной не более σ. Узорный объём найденных интервальных формальных понятий будет определять кластеры, то есть группы состояний с похожими распределениями достоверности в строках длиной $k * s$, и, следовательно, с похожими распределениями в каждой нечёткой матрице переходов $T(a_i)$. Все состояния найденного кластера объединяются в одно состояние, что позволяет минимизировать нечёткий автомат.

4. Поведение нечёткого автомата при минимизации числа состояний

При уменьшении числа состояний нечёткого автомата важно, чтобы поведение минимизированного нечёткого автомата соответствовало поведению исходного. Рассмотрим случай, в котором у нечёткой матрицы переходов автономного автомата главная диагональ состоит только из единиц. Очевидно, что если такая матрица допускает возможность кластеризации, то у матрицы с уменьшенным числом состояний главная диагональ также будет содержать максимальные значения.

Запись $X \geq Y$ означает, что $x_{ij} \geq y_{ij}$ для всех пар индексов (i, j). Напомним, что под операцией умножения нечётких матриц понимается min-max умножение, а именно $X \cdot Y = Z$, $z_{ij} = \max\limits_k \min\{x_{ik}, y_{kj}\}$.

Лемма 1. *Если у матрицы X на главной диагонали расположены единицы, то $X \cdot Y \geq Y$ и $Y \cdot X \geq Y$ для любой нечёткой матрицы Y.*

Доказательство. Пусть $Z = X \cdot Y$. По определению, $z_{ij} = \max\limits_k \min\{x_{ik}, y_{kj}\}$. При $k = i$ выполняется равенство $x_{ii} y_{ij} = \min\{x_{ii}, y_{ij}\} = \min\{1, y_{ij}\} = y_{ij}$, откуда $z_{ij} = \max\limits_k \min\{x_{ik}, y_{kj}\} \geq \min\{x_{ii}, y_{ij}\} = y_{ij}$.

Неравенство $Y \cdot X \geq Y$ доказывается аналогично. □

Лемма 2. *Для любого $s \geq 1$ имеют место соотношения*

$$X^{s+1} \geq X^s, \qquad X^s \geq X, \tag{1}$$

где X — нечёткая матрица с единичной главной диагональю. Поскольку элементы матриц X^s принадлежат конечному набору значений, соотношения (1) гарантируют, что, начиная с некоторого $n \in \mathbb{N}$ степени матрицы X стабилизируются: $X^n = X^{n+1} = \dots$.

Из этого утверждения с учётом того факта, что после кластеризации матрицы с единичной главной диагональю матрица с уменьшенным числом состояний также имеет единичную главную диагональ, вытекает важная теорема:

Теорема 1. Начиная с некоторого номера $m \in \mathbb{N}$ степени матрицы C с уменьшенным числом состояний также стабилизируются, то есть выполняется соотношение $C^m = C^{m+1}$.

Это означает, что поведение автономного нечёткого автомата с уменьшенным числом состояний совпадает с поведением исходного автомата.

Далее сформулируем теорему, относящуюся к произвольному нечёткому автомату:

Теорема 2. Если для любого символа $a \in A$ матрица нечёткого перехода $T(a)$ имеет единичную диагональ, то для любой бесконечной последовательности $P = \rho_1\rho_2 \dots \rho_m\rho_{m+1} \dots$ символов из алфавита A найдётся число $m \in \mathbb{N}$, начиная с которого поведение нечёткого автомата стабилизируется, т. е. $R_{m+1} = R_m$, где $R_{i+1} = R_i \cdot T(\rho_{i+1})$, $i \geq 1$, $R_1 = T(\rho_1)$.

Доказательство. Так как для любого символа $\rho_i \in A$ нечёткая матрица перехода $T(\rho_i)$ имеет единичную диагональ и $R_{i+1} = R_i \cdot T(\rho_{i+1})$, то в силу леммы 1 верно $R_{i+1} \geq R_i$. Таким образом, последовательность R_i увеличивается и в силу конечности набора значений нечёткой матрицы найдется номер $m \in \mathbb{N}$, такой, что начиная с m выполняется равенство $R_{m+1} = R_m$. $\square$

5. Распознавание нечёткого языка при минимизации числа состояний нечёткого автомата

Нечёткий автомат $F = (A, Q, \varphi, P_I)$ с помощью множества P_F распознаёт нечёткий язык $L(F)$ и для каждого распознаваемого слова $w \in L(F)$ определена достоверность $L(F)(w) = \max\limits_j(\min(\mu(q_{j_0}), \mu(q_{j_n}), M^j(w)))$. Достоверность $L(F')(w)$ слова w в новом автомате $F' = (A, Q', \varphi', P'_I)$, полученном при минимизации числа состояний автомата F с помощью интервальных формальных понятий, связана с $L(F)(w)$ следующим образом:

Теорема 3. *Достоверность $L(F)(w)$ слова $w \in L(F)$ не уменьшается при минимизации числа состояний нечёткого автомата F.*

Доказательство. Рассмотрим произвольное непустое слово $w \in L(F)$, $w = \omega_1\omega_2 \ldots \omega_n$ и достоверность слова w равна $L(F)(w)$.

При прочтении автоматом слова $w = \omega_1\omega_2 \ldots \omega_n$ образуется j-я траектория, которая описывается цепочкой изменений достоверностей и цепочкой изменений состояний:

$$q_{j_0} \longrightarrow q_{j_1} \longrightarrow \ldots \longrightarrow q_{j_{n-1}} \longrightarrow q_{j_n},$$

где q_{j_0} — начальное состояние для траектории j, а q_{j_n} — состояние автомата, в которое он попадает после прочтения слова w, проходя по траектории $j \in \Omega$.

Выберем два произвольных соседних состояния q_{j_i} и $q_{j_{i+1}}$ на этой траектории. То, что состояния являются соседними, означает, что при прочтении символа ω_{i+1} автомат F переходит из состояния q_{j_i} в состояние $q_{j_{i+1}}$ с достоверностью $\mu_{\omega_{i+1}}(q_{j_i}, q_{j_{i+1}})$.

При кластеризации множества состояний автомата $F = (A, Q, \varphi, P_I)$ мы получим новый автомат $F' = (A, Q', \varphi', P'_I)$. Множество состояний Q является объединением непересекающихся подмножеств: $Q = \bigcup\limits_i Q_i$. Каждое из подмножеств Q_i определяет состояние q'_i в новом нечётком автомате F'.

Как указано в разделе 3, вектор начального состояния P'_I в новом нечётком автомате выбирается так, чтобы достоверность $\mu(q'_{j_0})$ была максимальной среди всех достоверностей состояний, попавших в один кластер; таким же образом выбираются значения нового принимающего множества P'_F (нечёткого вектор-столбца).

Для соседних на j-й траектории состояний q_{j_i} и $q_{j_{i+1}}$ существуют два варианта: или они попали после кластеризации в один кластер (новое состояние q'_λ автомата F'), или в разные кластеры (разные состояния q'_ν и q'_ζ). В первом случае, как указано в разделе 3, в качестве диагонального элемента новой нечёткой матрицы переходов $T'(\omega_{i+1})$ выбирается значение

$$\mu_{\omega_{i+1}}(q'_\lambda, q'_\lambda) = \max_{q_m, q_n \in Q_\lambda} \mu_{\omega_{i+1}}(q_m, q_n).$$

то есть максимальная среди всех достоверностей перехода из состояния q_m в q_n исходной нечёткой матрицы переходов $T(\omega_{i+1})$, где q_m и q_n попали в один кластер $Q_\lambda = q'_\lambda$. Очевидно, что для q_{j_i} и $q_{j_{i+1}}$ верно неравенство

$$\mu_{\omega_{i+1}}(q_{j_i}, q_{j_{i+1}}) \leq \max_{q_m, q_n \in Q_\lambda} \mu_{\omega_{i+1}}(q_m, q_n).$$

Аналогично, если состояния q_{j_i} и $q_{j_{i+1}}$ попали в разные кластеры q'_ν и q'_ζ соответственно, достоверность перехода из состояния q_{j_i} в состояние $q_{j_{i+1}}$ при прочтении символа ω_{i+1} равна

$$\mu_{\omega_{i+1}}(q_{j_i}, q_{j_{i+1}}) \leq \max_{q_m \in Q_\nu, q_n \in Q_\zeta} \mu_{\omega_{i+1}}(q_m, q_n),$$

где $Q_\nu = q'_\nu$ и $Q_\zeta = q'_\zeta$.

Согласно определению, достоверность слова w, определяемая нечётким автоматом F' с помощью принимающего множества P'_F, вычисляется по формуле: $L(F')(w) = \max_j(\min(\mu(q'_{j0}), \mu(q'_{j_n}), M'^j(w)))$, где

$$M'^j(w) = \min_{0 \leq l \leq n-1} \mu_{\omega_l}(q'_{j_l}, q'_{j_{l+1}}).$$

Очевидно, что $M'^j(w) \geq M^j(w)$. Отсюда, с учётом неравенств $\mu(q'_{j0}) \geq \mu(q_{j0})$ и $\mu(q'_{j_n}) \geq \mu(q_{j_n})$, для j-й траектории имеем:

$$\min(\mu(q'_{j0}), \mu(q'_{j_n}), M'^j(w)) \geq \min(\mu(q_{j0}), \mu(q_{j_n}), M^j(w)).$$

Тогда, используя определение достоверности, получаем искомое неравенство:

$$L(F')(w) \geq L(F)(w). \qquad \square$$

Из теоремы вытекает утверждение о размере подъязыка $L^\beta(F)$:

Утверждение. Пусть $L^\beta(F)$ — язык, распознаваемый автоматом F с помощью множества P_F и содержащий слова с достоверностью не менее β, $L^\beta(F')$ — язык, распознаваемый минимизированным автоматом F' с помощью множества P'_F и содержащий слова с достоверностью не менее β. Тогда для любого β верно $L^\beta(F) \subseteq L^\beta(F')$.

6. Заключение

В настоящей работе рассматривается связь между задачей минимизации количества состояний нечёткого автомата и проблемой поиска интервальных формальных понятий с максимальным объёмом. Для исходного автомата $F = (A, Q, \varphi, P_I)$ для каждого входного символа из алфавита A определена матрица нечёткого перехода $T(A)$. Для каждого состояния из Q строка в нечёткой матрице показывает распределение достоверностей, с которыми автомат переходит во все состояния $\{q_1, q_2, \ldots, q_k\}$.

Если объединить состояния, имеющие похожие строки распределений, то можно получить новый автомат F' с числом состояний, меньшим, чем число состояний исходного автомата. Матрицу нечёткого перехода можно рассматривать как нечёткий контекст и применять методы теории формальных понятий. Контекст представляет собой узорную структуру, в которой объектам соответствуют состояния, а сложному описанию объектов — строки распределения достоверностей в нечёткой матрице. Нахождение строк контекста с похожими описаниями является задачей поиска интервальных формальных понятий с максимальным объёмом и содержанием, ширина которого по каждой компоненте ограничена значением σ. Существуют различные алгоритмы поиска интервальных понятий, а для больших размерностей нечёткого контекста можно использовать приближённый алгоритм.

Хорошим критерием того, что получаемый в результате кластеризации автомат наследует свойства исходного автомата, является схожее поведение их матриц нечёткого перехода при возведении в последовательные степени. В статье показано, что для случая нечётких матриц с единицами на главной диагонали как исходная, так и кластеризованная матрица автономного автомата обладают схожим свойством: последовательность их степеней является монотонно неубывающей и стабилизируется на некотором конечном шаге. Если входной алфавит содержит более одного символа и для каждого из них матрица нечёткого перехода обладает таким же свойством на диагонали, то результат последовательного умножения различных матриц перехода стабилизируется на некотором конечном шаге; это также верно для матриц с уменьшенным числом состояний.

Проблема уменьшения числа состояний нечёткого автомата порождает следующую интересную задачу: каким образом изменяется распознаваемый нечётким автоматом нечёткий язык? В работе доказано, что достоверность любого слова исходного нечёткого языка не уменьшается при кластеризации числа состояний нечёткого автомата. Этот факт позволяет заметить, что нечёткий язык, у которого каждое слово обладает достоверностью не меньше заданного параметра β исходного нечёткого автомата, является подмножеством языка с достоверностью слов не меньше β кластеризованного нечёткого автомата.

Дальнейшие исследования будут направлены на анализ свойств нечётких матриц, которые позволяют оценить, насколько и каким образом изменяется нечёткий язык при минимизации числа состояний нечёткого автомата с помощью интервальных формальных понятий.

Список литературы

[1] L. Zadeh, “Fuzzy Sets”, *Information and control*, **8** (1965), 338–353.

[2] W. G. Wee, K. S. Fu, “A Formulation of Fuzzy Automata and its Application as a Model of Learning Systems”, *IEEE Transactions on Systems Science and Cybernetics*, **5**:3 (July 1969), 215–223.

[3] B. Ganter, R. Wille, *Formal Concept Analysis*, Springer-Verlag, Berlin, 1999.

[4] B. Ganter, S. Kuznetsov, *Pattern Structures and Their Projections*, Preprint MATH-AL-14-2000, Technische Universit at Dresden, Herausgeber, Der Rektor, November, 2000.

[5] A. V. Galatenko, S. A. Nersisyan, D. N. Zhuk, “NP-Hardness of the Problem of Optimal Box Positioning”, *Mathematics*, 2019, № 7, 711.

[6] S. A. Nersisyan, V. V. Pankratieva, V. M. Staroverov, V. E. Podolskii, “A Greedy Clustering Algorithm Based on Interval Pattern Concepts and the Problem of Optimal Box Positioning”, *Journal of Applied Mathematics*, 2017, 4323590.

[7] А. А. Максимов, “Исследование сложных информационных систем с использованием универсально-алгебраических конструкций нечетких автоматов”, *Вестник Саратовского государственного социально-экономического университета*, 2006, № 14(3), 126–128.

[8] G. Birkhoff, *Lattice Theory*, American Mathematical Society colloquium publications, **25**, Amer. Math. Soc., 1940.

[9] R. Bělohlávek, “Determinism and fuzzy automata”, *Inf. Sci.*, **143**:1–4 (2002), 205–209.

[10] M. Kaytoue, S. Duplessis, S. O. Kuznetsov, A. Napoli, “Two FCA-Based methods for mining gene expression data”, *Proc. 7th International Conference on Formal Concept Analysis (ICFCA’2009), LNAI 5548*, eds. S. Ferre, S. Rudoplh, Springer-Verlag, Berlin, Heidelberg, 2009, 251–266.

[11] S. O. Kuznetsov, S. A. Obiedkov, “Comparing performance of algorithms for generating concept lattices”, *J. Expt. Theor. Artif. Intell.*, **14**:2–3 (2002), 189–216.

[12] J. Barretina, G. Caponigro, N. Stransky et al., “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”, *Nature*, **483** (2012), 603–607.

Minimization of the Number of States of a Fuzzy Automaton Using Interval Pattern Concepts

Pankratieva V.V.

Relationship between the problem of minimization of the number of states of fuzzy automata and the problem of mining of interval pattern concepts with maximum extent is considered.

The clustering method based on interval pattern concept mining combines states of a fuzzy automaton into subsets with similar patterns of confidence of transition into other states, where patterns are considered close to each other if the distance between them is bounded by a predetermined parameter σ. It is shown that for a certain type of fuzzy transition matrices the behavior of the original fuzzy automaton, as well as the behavior of the minimized one, eventually stabilizes. Moreover, it is proved that the membership value of each word recognized by the fuzzy automaton does not decrease after minimization. This fact allows comparing the fuzzy language recognized by the original automaton and the language recognized by the minimized automaton.

Keywords: fuzzy automata, interval pattern concepts, fuzzy languages.