ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 30 МАТЕМАТИКИ Выпуск 1
2023

А. В. Зязин, С. Ю. Мельников (Москва, МИРЭА, РУДН).
О возможности идентификации конечного автомата по его входным и выходным последовательностям в случае искажений.

УДК 519.713.4 DOI https://doi.org/10.52513/08698325_2023_30_1_1

Резюме: Предложен метод проверки того, может ли пара наблюдаемых искаженных последовательностей соответствовать истинным входным и выходным последовательностям эталонного двоичного автомата при условии, что начальное состояние автомата неизвестно, а доля удалений, вставок и замен символов невелика. Метод использует многоугольники, характеризующие статистические свойства автомата.

Ключевые слова: конечный автомат, идентификация автомата, эксперименты с автоматами.

1. Постановка задачи

Пусть $A = (X, Y, Q, h, f)$ — конечный сильносвязный автомат Мили, где $X = Y = \{0, 1\}$ — входной и выходной алфавиты, Q — множество состояний; $h\, Q \times X \to Q$ — функция переходов; $f\colon Q \times X \to Y$ — функция выходов. Пусть автомат A, начиная работать из начального состояния $q_0 \in Q$, перерабатывает последовательность $\chi^{(N)} = (x^{(1)}, x^{(2)}, \ldots, x^{(N)})$ в последовательность $\gamma^{(N)} = (y^{(1)}, y^{(2)}, \ldots, y^{(N)})$, где $x^{(i)}, y^{(i)} \in \{0, 1\}$, $1 \leqslant i \leqslant N$, $N \geqslant 1$. Результат работы [1] для случая, когда рассматриваются только частоты знаков во входной и выходной последовательности, заключается в следующем. Автомату A соответствует многоугольник R_A, расположенный в квадрате $[0, 1] \times [0, 1]$,

$$R_A = \mathrm{Conv}\left\{\left(\frac{c_x}{l(c)}, \frac{c_y}{l(c)}\right),\ c \in C_A\right\} \tag{1}$$

где C_A — множество элементарных циклов в графе переходов автомата, c_x, c_y — количество единиц во входной и выходной разметках цикла c, $l(c)$ — длина цикла c, $c \in C_A$, Conv — выпуклая оболочка множества точек на плоскости.

Для точки $z^{(N)} = \frac{1}{N}(\sum_{i=1}^{N} x^{(i)}, \sum_{i=1}^{N} y^{(i)})$ справедливо неравенство

$$\rho(z, R_A) \leqslant \frac{D}{N + D},$$

где D — диаметр автомата A, ρ — расстояние Чебышева на плоскости, $\rho((x, y), (x', y')) = \max\{|x - x'|, |y - y'|\}$.

Предположим, что наблюдаются не истинные последовательности $\chi^{(N)}$ и $\gamma^{(N)}$, а результаты исказивших их преобразований. Цель нашей работы — предложить способ проверки того, могут ли наблюдаемые искаженные последовательности соответствовать истинным для эталонного автомата, при условии, когда начальное состояние автомата неизвестно, а доля удалений, вставок, и замен знаков невелика.

2. Рассматриваемые преобразования последовательностей

2.1. Удаление символов. Пусть $\Omega_N = \{1, 2, \ldots, N\}$ — множество первых N натуральных чисел. Пусть заданы два подмножества $D_\chi, D_\gamma \subset \Omega_N$, состоящие из d_χ и d_γ элементов, соответственно, $0 \leqslant d_\chi, d_\gamma \leqslant N-1$. Из последовательности $\chi^{(N)} = (x^{(1)}, x^{(2)}, \ldots, x^{(N)})$ удалим символы, расположенные на тех местах, номера которых принадлежат D_χ, а из последовательности $\gamma^{(N)} = (y^{(1)}, y^{(2)}, \ldots, y^{(N)})$ удалим символы, номера которых принадлежат D_γ. Полученные после удаления последовательности обозначим соответственно χ_D и γ_D. Длины этих последовательностей равны $N - d_\chi$ и $N - d_\gamma$.

2.2. Вставка символов. Пусть в последовательности $\chi^{(N)}$ и $\gamma^{(N)}$ вставлены i_χ и i_γ двоичных символов, соответственно, $i_\chi, i_\gamma = 0, 1, \ldots$. Множества индексов вставленных элементов обозначим $I_\chi \subseteq \Omega_{N+i_\chi}$ и $I_\gamma \subseteq \Omega_{N+i_\gamma}$. Последовательности со вставленными элементами обозначим соответственно χ_i и γ_i. Длины этих последовательностей равны $N + i_\chi$ и $N + i_\gamma$.

2.3. Замена символов. Пусть заданы два подмножества $R_\chi, R_\gamma \subseteq \Omega_N$, состоящие из r_χ и r_γ элементов, соответственно, $0 \leqslant r_\chi, r_\gamma \leqslant N$. В последовательности $\chi^{(N)}$ выделим символы, расположенные на тех местах, номера которых принадлежат R_χ, а из последовательности $\gamma^{(N)}$ выделим символы, номера которых принадлежат R_γ. Осуществим замены выделенных символов на другие двоичные символы. Полученные после замен последовательности обозначим соответственно χ_R и γ_R. Длины этих последовательностей равны N.

3. Основной результат

Для конечной двоичной последовательности ε через $\|\varepsilon\|$ будем обозначать сумму ее элементов.

Лемма. *Справедливы неравенства*

$$\left|\frac{\|\chi^{(N)}\|}{N} - \frac{\|\chi_D\|}{N-d_\chi}\right| \leqslant \frac{d_\chi}{N-d_\chi}, \qquad \left|\frac{\|\gamma^{(N)}\|}{N} - \frac{\|\gamma_D\|}{N-d_\gamma}\right| \leqslant \frac{d_\gamma}{N-d_\gamma},$$
$$\left|\frac{\|\chi^{(N)}\|}{N} - \frac{\|\chi_I\|}{N+i_\chi}\right| \leqslant \frac{i_\chi}{N+i_\chi}, \qquad \left|\frac{\|\gamma^{(N)}\|}{N} - \frac{\|\gamma_I\|}{N+i_\gamma}\right| \leqslant \frac{i_\gamma}{N+i_\gamma},$$
$$\left|\frac{\|\chi^{(N)}\|}{N} - \frac{\|\chi_R\|}{N}\right| \leqslant \frac{r_\chi}{N}, \qquad \left|\frac{\|\gamma^{(N)}\|}{N} - \frac{\|\gamma_R\|}{N}\right| \leqslant \frac{r_\gamma}{N}.$$

Пусть:

$$z_D = \left(\frac{\|\chi_D\|}{N-d_\chi}, \frac{\|\gamma_D\|}{N-d_\gamma}\right), \quad z_I = \left(\frac{\|\chi_I\|}{N+i_\chi}, \frac{\|\gamma_I\|}{N+i_\gamma}\right), \quad z_R = \left(\frac{\|\chi_R\|}{N}, \frac{\|\gamma_R\|}{N}\right)$$

— точки на плоскости, построенные по наблюдаемым искаженным последовательностям, расстояние между точкой z *и множеством* R *определяется как* $\rho(z, R) = \inf\{\rho(z, z'),\ z' \in R\}$.

Теорема. *Пусть* R_A *— многоугольник автомата* A, *заданный соотношением* (1). *Справедливы неравенства*

$$\rho(z_D, R_A) \leqslant \frac{\max\{d_\chi, d_\gamma\}}{N - \max\{d_\chi, d_\gamma\}} + \frac{D}{N+D},$$
$$\rho(z_I, R_A) \leqslant \frac{\max\{i_\chi, i_\gamma\}}{N + \max\{i_\chi, i_\gamma\}} + \frac{D}{N+D},$$
$$\rho(z_R, R_A) \leqslant \frac{\max\{r_\chi, r_\gamma\}}{N} + \frac{D}{N+D}.$$

Сформулированная теорема дает метод проверки того, могут ли наблюдаемые искаженные последовательности соответствовать истинным для эталонного автомата в предположении, что начальное состояние автомата неизвестно. Метод состоит из двух этапов. На первом этапе строится многоугольник R_A эталонного автомата. На втором этапе строится одна из точек

$$z_D = \left(\frac{\|\chi_D\|}{N - d_\chi}, \frac{\|\gamma_D\|}{N - d_\gamma} \right), \quad z_I = \left(\frac{\|\chi_I\|}{N + i_\chi}, \frac{\|\gamma_I\|}{N + i_\gamma} \right) \quad \text{или} \quad z_R = \left(\frac{\|\chi_R\|}{N}, \frac{\|\gamma_R\|}{N} \right)$$

в соответствии с ожидаемым видом искажения. Далее проверяется выполнение соответствующего неравенства теоремы. Если неравенство нарушается, то наблюдаемые последовательности не соответствуют ни одному возможному варианту истинных последовательностей эталонного автомата ни для одного из его начальных состояний. Если неравенство выполнено, то выводов сделать нельзя.

Проанализируем вычислительную сложность предложенного метода, если число $|Q|$ состояний автомата велико. Она определяется двумя слагаемыми: сложностью предварительного построения многоугольника и сложностью проверки неравенств теоремы. Из результатов [1] следует, что сложность предварительного построения многоугольника ограничена величиной $O(|Q|2^{|Q|})$, хотя для некоторых классов автоматов ([2]) возможны аналитические методы построения многоугольника.

Если многоугольник R_A уже построен, то сложность проверки неравенств теоремы не более чем в 4 раза превышает сложность проверки принадлежности заданной точки выпуклому многоугольнику R_A. Вычислительную сложность последней задачи можно оценить ([3]) величиной $O(\mathrm{Log}\,\nu)$, где ν — число вершин многоугольника. Для оценки ν воспользуемся тем, что все вершины многоугольника R_A имеют вид $\left(\frac{p_1}{q_1}, \frac{p_2}{q_2}\right)$, $0 \leqslant p_i \leqslant q_i \leqslant |Q|$, $i = 1, 2$. Оценив возможное число различных абсцисс вершин многоугольника, в силу выпуклости многоугольника, получим $\nu \leqslant |Q|^2$. Поэтому сложность проверки неравенств теоремы в случае предварительно построенного многоугольника оценивается как $O(\mathrm{Log}\,|Q|)$.

Информативность метода многогранников, связанная с условиями выполнения неравенства $\rho(z_R, R_A) \leqslant \frac{D}{N+D}$, изучалась в [4]. В сформулированной теореме неравенства имеют вид $\rho(z_R, R_A) \leqslant \alpha + \frac{D}{N+D}$, где величина $\alpha \geqslant 0$ растет с увеличением доли искажений. Поэтому информативность предложенного метода уменьшается с увеличением доли искажений.

4. Выводы.

Предложено обобщение метода идентификации автоматов на основе многоугольников по частотным характеристикам знаков входной и выходной последовательностей. Обобщение проведено для случая, когда эти последовательности подвергаются трем видам искажений: вставка, удаление, замена символов. Показано, что для искаженных последовательностей метод идентификации автоматов на основе многоугольников остается работоспособным. Однако с увеличением доли искажений информативность метода снижается. Вычислительная сложность метода практически не зависит от уровня искажений.

СПИСОК ЛИТЕРАТУРЫ

1. *Melnikov S. Yu.*, *Samouylov K. E.* Polyhedra of finite state machines and their use in the identification problem. In: Internet of Things, Smart Spaces, and Next Generation Networks and Systems. 20th International Conference, NEW2AN 2020, and 13th Conference, ruSMART. (Saint Petersburg, August 26–28, 2020.) Proceedings. Part II. / Ed. by O. Galinina, S. Andreev, S. Balandin, Ye. Koucheryavy. Cham etc.: Springer, 2020, p. 110–121. (Ser. Lect. Notes Comput. Sci. V. 12526.)
2. *Melnikov S. Yu.*, *Samouylov K. E.* Polygons characterizing the joint statistical properties of the input and output sequences of the binary shift register. — Proceedings of the 4th International Conference on Future Networks and Distributed Systems. 2020, p. 1–6.

3. *Preparata F. P.*, *Shamos M. I.* Computational Geometry - An Introduction. Springer-Verlag, 1985, 398 p.
4. *Melnikov S. Yu.*, *Samouylov K. E.*, *Zyazin A. V.* Estimating a Polyhedron Method Informativeness in the Problem of Checking the Automaton by the Statistical Properties of the Input and Output Sequences. — International Conference on Distributed Computer and Communication Networks. Cham : Springer Nature Switzerland, 2022, p. 42–51.

UDC 519.713.4 DOI https://doi.org/10.52513/08698325_2023_30_1_1

Zyazin A. V., ***Melnikov S. Yu.*** (Moscow, MIREA — Russian Technological University, Moscow, Peoples' Friendship University of Russia named after Patrice Lumumba). **On the possibility of identifying a finite automaton by its input and output sequences in case of distortions.**

Abstract: A method is proposed for checking whether a pair of observed distorted sequences can correspond to the true input and output sequences of a reference binary automaton, provided that the initial state of the automaton is unknown, and the proportion of deletions, insertions, and replacements of symbols is small. The method uses polygons that characterize the statistical properties of the automaton.

Keywords: finite automaton, identification of automata, experiments with automata.