Критерий почти полного прогнозирования сверхслова в многозначном алфавите

И. К. Ведерников[1]

Автомат прогнозирует следующий символ входного сверхслова, если он выдает этот символ на выходе в предыдущий момент времени.

В работе для произвольного сверхслова в многозначном алфавите исследуется вопрос его почти полного прогнозирования, т.е. когда прогноз правильный почти всегда. Получен результат, позволяющий сузить класс автоматов, которыми прогнозируется сверхслово, а также доказан критерий почти полной прогнозируемости сверхслова.

Ключевые слова: почти полное прогнозирование, прогнозирующий автомат, автоматное прогнозирование сверхслов, критерий прогнозируемости.

1. Введение

В статье А.Г. Вереникина и Э.Э. Гасанова [1] были введены прогнозирующие автоматы — конечные автоматы, предсказывающие сверхслово или множество сверхслов. Говорят, что автомат прогнозирует сверхслово, если через некоторое конечное время после начала, он начинает в момент времени t выдавать элемент входной последовательности под номером $t+1$.

Оказалось, что полностью прогнозируемы только периодические сверхслова, изначально это было доказано для двоичного алфавита, но в работе [2] данный результат был обобщен на случай k-значных логик.

В работе [3] А.А. Мастихиной было введено понятие частичного прогнозирования, которое имеет место в случае, когда автомат угадывает следующий символ не обязательно в каждый момент времени, но достаточно часто. В одной из следующих работ А.А. Мастихина [4] предъявила критерий частичной прогнозируемости общерегулярных сверхсобытий в двоичном алфавите. Также вопрос частичного прогнозирования сверхсобытий исследовался в работе [5].

[1] *Ведерников Илья Константинович* — аспирант каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: ikvedernikov@gmail.com.

Vedernikov Iliya Konstantinovich — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

Впервые вопрос прогнозируемости одного сверхслова поднимался А.А. Мастихиной в работе [3], в частности, было доказано, что существует сверхслово не прогнозируемое ни одним автоматом. Однако в поздних работах А.А. Мастихиной акцент делался исключительно на сверхсобытия. В данной работе же исследуется вопрос прогнозирования для одного сверхслова в многозначном алфавите, причем не обязательно общерегулярного. Было введено понятие почти полного прогнозирования – это означает, что автомат угадывает сверхслово почти всегда, т.е. степень частичного прогнозирования равна единице. В результате получено, что наилучшая степень прогнозирования достигается в определенном классе автоматов, а также получен критерий почти полной прогнозируемости одного сверхслова.

Автор выражает благодарность профессору Э.Э. Гасанову за постановку задачи и помощь в работе.

2. Основные понятия и формулировка результата

Введем основные определения.

Пусть $E_k = \{0, 1, \dots, k-1\}$ – конечный алфавит. Через E_k^* и E_k^∞ обозначим соответственно множество всех слов конечной длины и множество всех сверхслов в алфавите E_k. По определению будем считать, что пустое слово Λ принадлежит E_k^*. Подмножества E_k^* называются *событиями*, а подмножества E_k^∞ — *сверхсобытиями.*

Длину слова α обозначим $|\alpha|$, по определению $|\Lambda| = 0$. Если α — сверхслово, то $|\alpha| = \infty$.

Если α — сверхслово в алфавите E_k, n — натуральное число, то n-ую букву сверхслова α будем обозначать $\alpha(n)$, а через $\alpha\underset{n}{]}$ обозначим префикс длины n сверхслова α, т.е. $\alpha\underset{n}{]} = \alpha(1)\alpha(2)\dots\alpha(n)$.

Если α — слово в алфавите E_k, $|\alpha| = m$, n — натуральное число, $n < m$, то через $\underset{n}{[}\alpha$ обозначим суффикс длины n слова α, т.е. $\underset{n}{[}\alpha = \alpha(m-n+1)\dots\alpha(m)$. Если $n = 0$, то положим $\underset{n}{[}\alpha = \Lambda$.

В работе рассматриваются конечные инициальные автоматы в соответствии с нотацией из [6]:

$$V = (E_k, Q, E_k, \varphi, \psi, q_0),$$

где $E_k = \{0, 1, ..., k-1\}$ – входной и выходной алфавит, Q – множество состояний, которое является конечным подмножеством некоторого фиксированного счетного множества, $\varphi : Q \times E_k \to Q$ – функция переходов, $\psi : Q \times E_k \to E_k$ – функция выходов, q_0 – начальное состояние.

Если на вход инициальному автомату V подается слово или сверхслово $x = x(1)x(2)\dots$, на выходе получается слово или сверхслово $y = y(1)y(2)\dots$, и $q(t)$ означает состояние автомата в момент времени t, то функционирование автомата задается системой уравнений

$$\begin{cases} q(1) = q_0, \\ q(t+1) = \varphi\big(q(t), x(t)\big), \\ y(t) = \psi\big(q(t), x(t)\big), \end{cases}$$

где $t \in \mathbb{N}$.

Также определим автомат без выхода $V = (E_k, Q, \varphi, q_0)$, где E_k, Q, φ, q_0 определяются аналогично определению выше, а функционирование задается системой

$$\begin{cases} q(1) = q_0, \\ q(t+1) = \varphi\big(q(t), x(t)\big), \end{cases}$$

где $t \in \mathbb{N}$.

Функции φ и ψ естественно расширяются на $Q \times E_k^*$, а именно, если $\alpha \in E_k^*$, $a \in E_k$, то индуктивно определим

$$\varphi(q, \alpha a) = \varphi(\varphi(q, \alpha), a),$$

$$\psi(q, \alpha a) = \psi(\varphi(q, \alpha), a).$$

Введем также обозначения

$$\overline{\varphi}(q, \alpha) = \varphi(q, \alpha]_1)\varphi(q, \alpha]_2)\dots\varphi(q, \alpha),$$

$$\overline{\psi}(q, \alpha) = \psi(q, \alpha]_1)\psi(q, \alpha]_2)\dots\psi(q, \alpha),$$

если α — слово, если же α — сверхслово, то

$$\overline{\varphi}(q, \alpha) = \varphi(q, \alpha]_1)\varphi(q, \alpha]_2)\dots\varphi(q, \alpha]_n)\dots,$$

$$\overline{\psi}(q, \alpha) = \psi(q, \alpha]_1)\psi(q, \alpha]_2)\dots\psi(q, \alpha]_n)\dots.$$

Если α — сверхслово в алфавите A, то *пределом* сверхслова α назовем такое множество $A' \subseteq A$, что в сверхслове α бесконечное число раз встречаются символы из A' и только они. Этот факт будем обозначать $A' = \lim \alpha$.

Если есть событие R_1 и событие или сверхсобытие R_2, то через $R_1 R_2$ обозначим их *произведение*, то есть все слова (сверхслова) вида ab, где $a \in R_1, b \in R_2$ (ab – конкатенация слов a и b).

Если R – событие, то обозначим R^* – *итерация* события R, то есть $R^* = R \cup R^2 \cup R^3 \cup ... \cup R^i ...$, а R^∞ – *сверхитерация* события R, $R^\infty = \{a_1 a_2 a_3 ... |\ a_i \in R, i = 1, 2, 3...\}$.

Сверхсобытие R *представимо* автоматом $V = (E_k, Q, E_k, \varphi, \psi, q_0)$ с помощью семейства F, $F \subseteq 2^Q$, тогда и только тогда, когда для любого $\alpha \in R$, существует $Q' \in F$, такое, что $\lim \overline{\varphi}(q_0, \alpha) = Q'$.

Пусть $t \in \mathbb{N}$, скажем, что $(t+1)$-й символ сверхслова $\alpha = \alpha(1)\alpha(2)\ldots \alpha(t+1)\ldots$ или слова $\alpha = \alpha(1)\alpha(2)\ldots\alpha(t')$, $t' > t$, *угадан* автоматом $V = (A, Q, B, \varphi, \psi, q_0)$, если $\psi(q_0, \alpha]_t) = \alpha(t+1)$.

Пусть $\alpha \in E_k^\infty$, обозначим $\sigma_\alpha(V) = \underline{\lim}_{n\to\infty} N_n/n$, где N_n – количество угаданных автоматом V символов в слове $\alpha\underset{n}{]}$. Будем говорить, что $\sigma_\alpha(V)$ – *степень прогнозирования сверхслова* α автоматом V. Если $\alpha \in E_k^*$, $|\alpha| = n$, обозначим $\sigma_\alpha(V) = N/n$, где N – количество угаданных автоматом V символов в слове α. Будем говорить, что $\sigma_\alpha(V)$ – *степень прогнозирования слова* α автоматом V.

Считаем, что множество сверхслов R *частично прогнозируемо*, если существует такой автомат V, что степень прогнозирования для каждого сверхслова множества R строго больше нуля. Обозначим $\sigma_R(V) = \inf\limits_{\alpha\in R} \sigma_\alpha(V)$.

Если $\mathfrak{K}$ — некоторый класс автоматов, то степень прогнозирования сверхсобытия R на автоматах из этого класса определим как $\sigma_R(\mathfrak{K}) = \sup\limits_{V\in\mathfrak{K}} \sigma_R(V)$.

Будем говорить, что автомат $V = (E_k, Q, E_k, \varphi, \psi, q_0)$, прогнозирующий сверхсобытие R, лежит в классе представляющих относительно R, если существует автомат $V_0 = (E_k, Q, \varphi, q_0)$, представляющий R с помощью некоторого F, $F \subseteq 2^Q$.

Определим некоторый способ задания выходов, т.е. способ описания функции $\psi : Q \times E_k \to E_k$. Задавать выходную функцию будем, помечая ребра диаграммы Мура – в каждом состоянии отметим одно исходящее ребро. Тогда функция выхода будет описана следующим образом: если у нас отмечено ребро по символу b, $b \in E_k$, исходящее из некоторого состояния q', то для всех q и a, таких что $\varphi(q, a) = q'$, значение выходной функции $\psi(q, a)$ будет равно b. Понятно, что задавая таким образом функцию ψ для каждого состояния, в итоге мы полностью определим ее. Функцию выхода, полученную таким образом, будем называть *размеченной.*

Заметим, что угадывание происходит, когда выход в предыдущий момент времени равен входу в данный момент времени. Поэтому если автомат проходит по отмеченной стрелке, то угадывание происходит.

Обозначим $\mathfrak{A}$ – класс всех автоматов, $\mathfrak{R}(R)$ — множество автоматов, которые лежат в классе представляющих относительно R, $\mathfrak{M}$ – множество всех автоматов с размеченными функциями выхода, а $\mathfrak{MR}(R)$ — все автоматы из $\mathfrak{R}(R)$ с размеченной функцией выхода.

Будем говорить, что сверхслово α *почти полностью прогнозируемо*, если $\sigma_\alpha(\mathfrak{A}) = 1$.

Будем говорить, что H – *конечное приведенное множество слов* над E_k, если для любых α_1, α_2 из H выполнено, что α_1 не является подсловом α_2 и наоборот. Через $M(H)$ обозначим максимальную длину слова из H.

Представлением сверхслова β назовем последовательность слов $\mathfrak{B} = \{\alpha_i^{s_i}\}$ такую, что $\beta = \alpha_0^{s_0}\alpha_1^{s_1}\ldots\alpha_m^{s_m}\ldots$. Слово $\alpha_i^{s_i}$ будем называть *i-ым элементом* представления.

Накрытием сверхслова β назовем пару $(\mathfrak{B}, H)$, где $\mathfrak{B}$ – представление сверхслова β, H – конечное приведенное множество слов.

Далее определим некоторые характеристики накрытия:

1) Через $\mathfrak{C}_H(\mathfrak{B})$ обозначим *покрывающее множество* и определим его следующим алгоритмом (опишем i-й шаг): возьмем i-ый элемент $\mathfrak{B}$, если $\alpha_i \in H$, то добавим $\alpha_i^{s_i}$ в $\mathfrak{C}_H(\mathfrak{B})$, если нет – ничего не делаем. После этого перейдем к шагу $i+1$.
   Отметим, что $\mathfrak{C}_H(\mathfrak{B})$ – множество с повторениями, то есть может содержать два и более одинаковых слова.

2) $C(\mathfrak{C}_H(\mathfrak{B}), n)$ – количество элементов покрывающего множества, входящих в префикс сверхслова длины n.

3) $L(\mathfrak{C}_H(\mathfrak{B}), n)$ – суммарная длина всех элементов покрывающего множества, входящих в префикс сверхслова длины n.

Далее сформулируем основные результаты данной работы:

Теорема 1. *Пусть автомат $V = (E_k, Q, E_k, \varphi, \psi, q_0)$ прогнозирует сверхсобытие R со степенью σ, ψ – не размеченная, тогда существует автомат $\widehat{V} = (E_k, \widehat{Q}, E_k, \widehat{\varphi}, \widehat{\psi}, \widehat{q_0})$, где $\widehat{\psi}$ – размеченная, который прогнозирует R со степенью σ. Более того, если $V \in \mathfrak{R}(R)$, то $\widehat{V} \in \mathfrak{MR}(R)$.*

Теорема 2. *Пусть $\beta \in E_k^\infty$. Тогда β угадываемо со степенью 1 тогда и только тогда, когда существует накрытие $(\mathfrak{B}, H)$ такое, что выполнено*

$$\lim_{n\to\infty} \frac{L(\mathfrak{C}_H(\mathfrak{B}), n) - M(H)\cdot C(\mathfrak{C}_H(\mathfrak{B}), n)}{n} = 1.$$

Замечание 1. *Условие теоремы 2 эквивалентно следующему:*

$$\lim_{n\to\infty} \frac{L(\mathfrak{C}_H(\mathfrak{B}), n) - C(\mathfrak{C}_H(\mathfrak{B}), n)}{n} = 1.$$

3. Достаточность размеченных функций выхода.

Теорема 1. *Пусть автомат* $V = (E_k, Q, E_k, \varphi, \psi, q_0)$ *прогнозирует сверхсобытие* R *со степенью* σ, ψ *– не размеченная, тогда существует автомат* $\widehat{V} = (E_k, \widehat{Q}, E_k, \widehat{\varphi}, \widehat{\psi}, \widehat{q_0})$, *где* $\widehat{\psi}$ *– размеченная, который прогнозирует* R *со степенью* σ. *Более того, если* $V \in \mathfrak{R}(R)$, *то* $\widehat{V} \in \mathfrak{MR}(R)$.

Доказательство. Будем преобразовывать автомат V так, чтобы его функция выхода стала размеченной.

Рассмотрим произвольное состояние q. Пусть $Prev_q = \{(q_0, c_0), \ldots, (q_n, c_n)\}$ – множество пар состояние – символ таких, что в автомате V для любой пары $(q', c) \in Prev_q$ выполняется $\varphi(q', c) = q$.

Если в этом состоянии функция выхода автомата V удовлетворяет требованию размеченности, то есть существует символ b, что для любой пары $(q', c) \in Prev_q$ выполнено, что $\psi(q', c) = b$, то для состояния q условие размеченности выполнено, и мы переходим к рассмотрению другого состояния.

Если не удовлетворяет, то существуют $b_1, \ldots, b_l$, $l \leq n, l \leq k$, и представление множества $Prev_q = Prev_q^1 \cup \cdots \cup Prev_q^l$ такие, что для любой пары $(q', c) \in Prev_q^s, s \in \{1, \ldots, l\}$, выполнено, что $\psi(q', c) = b_s$.

Далее получим из V некоторый автомат $V' = (E_k, Q', E_k, \varphi', \psi', q_0')$:

1) Заменим состояние q на состояния $\{q^1, \ldots, q^l\}$, то есть $Q' = Q \cup \{q^1, \ldots, q^l\} \setminus \{q\}$.

2) Функции переходов и выходов определим следующим образом:

    а) Для любой пары (q', c) из $Prev_q^s$, $q' \neq q$, положим, что $\varphi'(q', c) = q^s$, а $\psi'(q', c) = b_s$.

    б) Пусть $j \in \{1, \ldots, l\}$, тогда для любого $a \in E_k$, если $\varphi(q, a) \neq q$, положим $\varphi'(q^j, a) = \varphi(q, a)$, $\psi'(q^j, a) = \psi(q, a)$. Если $\varphi(q, a) = q$, то пара (q, a) принадлежит некоторому $Prev_q^s$, тогда положим $\varphi'(q^j, a) = q^s$, $\psi'(q^j, a) = b_s$.

    в) Для остальных случаев положим $\varphi'(q', c) = \varphi(q', c)$, и $\psi'(q', c) = \psi(q', c)$.

3) Если q было начальным состоянием автомата V, то $q_0' = q^1$, иначе $q_0' = q_0$.

Заметим, что, исходя из построения, в любой момент времени t и для любого входного сверхслова α выполнено, что $\psi(q_0, \alpha|_t) = \psi'(q_0', \alpha|_t)$, то есть автомат V' неотличим от V. Следовательно, если автомат V угадывал некоторый символ $\alpha(m+1)$, то и автомат V' будет угадывать этот символ. Более того в автомате V' в состояних $q^1, \ldots, q^l$ функция выхода ψ' – размеченная, и, если в автомате V было r состояний с не размеченной функцией выхода, то в автомате V' их будет $r-1$.

Теперь рассмотрим ситуацию, когда $V \in \mathfrak{R}(R)$. Тогда существует автомат $V_0 = (E_k, Q, \varphi, q_0)$, который представляет R с помощью $F = \{Q_1, \ldots, Q_h\}$, $F \subseteq 2^Q$. После операции описанной выше, мы получили автомат $V' = (E_k, Q', E_k, \varphi', \psi', q_0')$ и далее построим множество F', с помощью которого автомат $V_0' = (E_k, Q', \varphi', q_0')$ представляет R. Для этого рассмотрим каждый элемент Q_i множества F и сделаем следующее:

1) Пусть $q \notin Q_i$, тогда добавим Q_i в F'.

2) Пусть $q \in Q_i$. Рассмотрим множества $S_{p,i} = L_p \cup Q_i \setminus \{q\}$, где $L_p \in 2^{\{q^1,\ldots,q^l\}} \setminus \{\varnothing\}$. Далее, если существует такое сверхслово $\alpha \in R$, что $\lim(\alpha) = S_{p,i}$, то добавим $S_{p,i}$ в F'.

Заметим, что если в автомате V_0 предел сверхслова был равен Q_i, то в автомате V_0' его предел будет равен либо множеству $S_{p,i}$ для некоторого p, либо самому Q_i. В результате получим, что если сверхслово $\alpha \in R$, то $\lim(\alpha) \in F'$. Обратное тоже верно, так как если $\lim(\alpha) = S_{p,i}$ или $\lim(\alpha) = Q_i$ в автомате V_0', то в автомате V_0 его предел равен Q_i, а значит $\alpha \in R$.

Тем самым мы показали, что в результате операции разделения состояний, сохраняется принадлежность к классу представляющих автоматов.

Далее возьмем V' и проделаем с ним операцию описанную выше, получим автомат V'' с $r-2$ состояниями, в которых функция выхода не размеченная. Проделав эту операцию r раз, получим автомат $\widehat{V}$, у которого функция выхода в каждом состоянии размеченная, и который по построению прогнозирует R со степенью σ. И так как сохраняется принадлежность к классу представляющих автоматов, то итоговый автомат $\widehat{V}$ будет принадлежать классу $\mathfrak{M}\mathfrak{R}(R)$, если автомат V был из $\mathfrak{R}(R)$.

$\square$

Замечание 2. *Из теоремы явно следует, что для любого сверхслова β выполнено $\sigma_\beta(\mathfrak{A}) = \sigma_\beta(\mathfrak{M})$. Поэтому далее будем рассматривать автоматы только из $\mathfrak{M}$.*

4. Доказательство вспомогательных утверждений и вспомогательные построения.

Введем несколько вспомогательных определений:

Сильно связным множеством назовем такое множество состояний C, $C \subseteq Q$, автомата $V = (E_k, Q, E_k, \varphi, \psi, q_0)$, что для любых $q', q'' \in C$ найдется такое слово α из E_k^*, что $\varphi(q', \alpha) = q''$ и $\overline{\varphi}(q', \alpha) \in C^*$, т.е. по слову α автомат переходит из состояний q' в состояние q'', проходя только состояния из C. Множество, состоящее из одного состояния, по определению считается сильно связным.

Любое сильно связное множество состояний C назовем *автоматным циклом*, если $|C| > 1$. Одноэлементное множество состояний $C = \{q\}$ является *автоматным циклом* только если существует символ a из E_k, что $\varphi(q, a) = q$. *Длиной автоматного цикла* назовем число состояний в автоматном цикле.

Пусть $V = (E_k, Q, E_k, \varphi, \psi, q_0)$ и ψ – размеченная. *Размеченным циклом* в автомате V назовем автоматный цикл автомата V, в котором все переходы по циклу отмечены.

Лемма 1. *Пусть у автомата V, $V \in \mathfrak{M}$, нет размеченного цикла, тогда для любого сверхслова β выполнено, что $\sigma_\beta(V) < 1$.*

Доказательство. Заметим, что функция ψ – размеченная, а значит угадывание происходит только тогда, когда автомат проходит по отмеченной стрелке в диаграмме Мура. По условию, в автомате нет размеченного цикла, следовательно, каждые m тактов, где m – количество состояний автомата V, будет не более $m-1$ угадывания. Отсюда следует, что $\sigma_\beta(V) \leq \frac{m-1}{m} < 1$.

□

Теперь пусть H – конечное приведенное множество слов над E_k, опишем некоторое частичное построение автомата, которое пригодится при доказательстве теоремы:

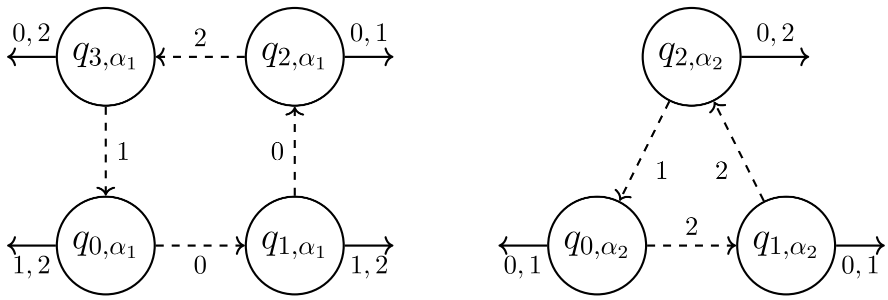

Рис. 1. Пример построения 1.

Построение 1. *Для каждого слова из H построим размеченный автоматный цикл.*

1) Пусть $\alpha \in H$, $\alpha = a_0 \dots a_{l-1}$, тогда в цикле будет l состояний, обозначим их $\{q_{0,\alpha}, \dots q_{l-1,\alpha}\}$.

2) Функцию переходов зададим следующим образом: $\varphi(q_{i,\alpha}, a_i) = q_{i+1,\alpha}$, если $i \in \{0, \dots, l-2\}$. Если $i = l-1$, то $\varphi(q_{l-1,\alpha}, a_{l-1}) = q_{0,\alpha}$. Остальные переходы оставим неопределенными.

3) Функцию выхода зададим с помощью разметки. В каждом состоянии $q_{i,\alpha}$ отметим переход a_i, $i \in \{0, \dots, l-1\}$.

Пример к построению 1. Пусть $H = \{\alpha_1, \alpha_2\}$, $H \subset E_3^*$, где $\alpha_1 = 0021$, $\alpha_2 = 221$. В результате построения получим два цикла, которые изображены на рис. 1, пунктиром обозначены отмеченные переходы.

Далее пусть есть автомат $V = (E_k, Q, \varphi, q)$, и каждому слову из H сопоставлено одно состояние из Q, причем разным словам – разные состояния. Будем говорить, что *автомат V разделяет вход по множеству H*, если для любого слова $\gamma\alpha$, где α из H, а γ такое, что ни одно слово из H не содержится как подслово в слове $\gamma\alpha\,]_{|\alpha|-1}$, выполнено $\varphi(q, \gamma\alpha) = q_\alpha$, где q_α – состояние, сопоставленное слову α. Данное определение означает, что в автомате выделено $|H|$ состояний, и как только во входном слове встречается подслово равное слову из H, то автомат попадает в соответствующее выделенное состояние и не выходит оттуда.

Отметим, что определение выше ставит задачу поиска первого попавшегося слова из некоторого фиксированного множества в качестве подслова. Подобную задачу уже давно рассматривают, и известным результатом в этой области является алгоритм Ахо-Корасик [7], который реализует поиск множества подстрок в некоторой данной строке. Алгоритм строит конечный автомат, которому затем передаёт строку поиска. В лемме 2 описывается построение подобного автомата, но с учетом того, что множество искомых слов приведенное и искать нужно до первого вхождения, эти условия гарантируют, что получится автомат, который разделяет вход по множеству H.

Лемма 2. Пусть H – конечное приведенное множество слов над E_k, тогда существует автомат V, разделяющий вход по H.

Доказательство. Пусть $H = \{\alpha_0, \dots, \alpha_{p-1}\}$, и пусть префиксное дерево для слов из H имеет m вершин, среди которых p листьев. Обозначим вершины как $U = \{v_0, \dots, v_{m-1}\}$, причем v_0 – корень дерева, а $\{v_{m-p}, \dots, v_{m-1}\}$ – листовые вершины.

Далее построим автомат V, разделяющий вход по H. Положим $Q = \{q_0, \dots, q_{m-1}\}$ – состояния автомата, при этом q_0 – начальное состояние. Каждому состоянию автомата q_i сопоставим вершину v_i из U и зададим функцию переходов:

1) Если в префиксном дереве есть ребро из вершины v_i в вершину v_j по символу a, то положим $\varphi(q_i, a) = q_j$.

2) Если q_i – состояние, которому соответствует листовая вершина, то $\varphi(q_i, a) = q_i$ для любого a из E_k.

3) Если в префиксном дереве из вершины v_0 не выходит ребра с символом a, то $\varphi(q_0, a) = q_0$.

4) Возьмем состояние $q_i,\ i \in \{1, \dots, m-p-1\}$, пусть $\delta,\ |\delta| = l$, – слово, которое проводит из корня v_0 в вершину v_i, и пусть в пункте 1 в состоянии q_i было задано $k - k'$ переходов, $k' < k$. Без ограничения общности можно предположить, что осталось задать переходы по символам $\{0, \dots, k' - 1\}$.

   Рассмотрим произвольный символ $h,\ h \in \{0, \dots, k' - 1\}$. Возьмем суффикс $[_{l-1}\,\delta$, если существует слово α из H такое, что $([_{l-1}\,\delta)h = \alpha\,]_l$, то положим $\varphi(q_i, h) = \varphi(q_0, \alpha\,]_l)$. Если не существует, то возьмем суффикс $[_{l-2}\,\delta$ и префиксы слов из H длины $l - 1$ и т.д.. Если же для любого $l',\ l' \in \{1, \dots, l\}$, и любого слова $\alpha,\ \alpha \in H$, верно, что $([_{l-l'}\,\delta)h \neq \alpha\,]_{l-l'+1}$, то положим $\varphi(q_i, h) = q_0$.

Заметим, что множество H приведенное, то есть ни одно слово не является подсловом другого, поэтому построение в пункте 4 корректно. Также заметим, что по построению для любого слова $\gamma,\ \gamma \in E_k^*$, выполнено, что как только первый раз встречается слово α из H в качестве подслова γ, то автомат заходит в состояние $\varphi(q_0, \alpha)$ и уже не выходит оттуда, при этом разным словам из H соответствуют разные состояния. Тем самым мы построили автомат V, который разделяет вход по множеству H.

□

Пример к лемме 2. Пусть $H = \{20, 221, 1110, 1121\},\ H \subset E_3^*$. В результате построения как в лемме 2 получим автомат, изображенный на рис. 2.

5. Доказательство критерия угадываемости.

Теорема 2. *Пусть $\beta \in E_k^\infty$. Тогда β угадываемо со степенью 1 тогда и только тогда, когда существует накрытие $(\mathfrak{B}, H)$ такое, что выполнено*

$$\varliminf_{n\to\infty} \frac{L(\mathfrak{C}_H(\mathfrak{B}), n) - M(H) \cdot C(\mathfrak{C}_H(\mathfrak{B}), n)}{n} = 1.$$

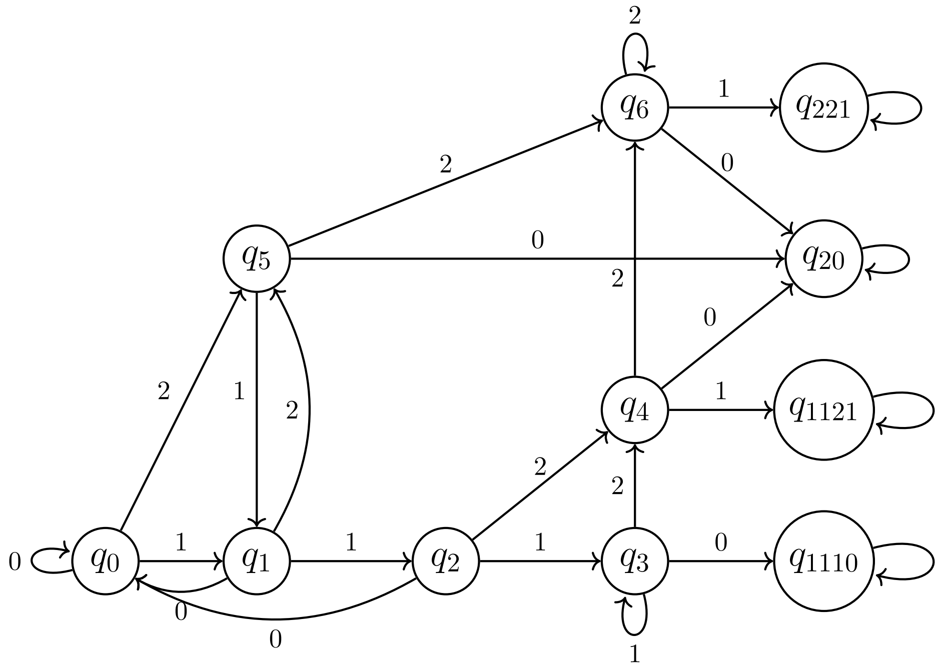

Рис. 2. Пример к лемме 2.

Доказательство. Необходимость. Пусть $\sigma_\beta(\mathfrak{A}) = 1$. Возьмем автомат $V = (E_k, Q, E_k, \varphi, \psi, q_0),\ |Q| = m$, такой, что $\sigma_\beta(V) = 1,\ \ V \in \mathfrak{M}$, по условию и теореме 1 такой существует.

По лемме 1 автомат V имеет хотя бы один размеченный цикл, пусть у него их p штук, обозначим их $\{C_0, \dots C_{p-1}\}$. Зафиксируем у каждого цикла C_i произвольное состояние, которое назовем начальным и обозначим $q_{0,i},\ i \in \{0, \dots, p-1\}$. Теперь пусть слово, которое проводит автомат от начального состояния цикла $q_{0,i}$ до него же – это соответственно α_i. Если среди этих слов есть такие, что $\alpha_i^\infty = \alpha_j^\infty$, то выбросим максимальное по длине, в результате получим множество слов $\{\alpha_0, \dots, \alpha_{p'-1}\},\ p' \le p$. Положим $H = \{\alpha_0^{s_0}, \dots, \alpha_{p'-1}^{s_{p'-1}}\}$, где $\{s_0, \dots, s_{p'-1}\}$ минимальные натуральные числа такие, что для любого i слово $\alpha_i^{s_i}$ не является подсловом $\alpha_j^{s_j},\ j \in \{0, \dots, i-1, i+1, \dots, p'-1\}$. По построению H – приведенное множество слов.

Представление $\mathfrak{B}$ построим по автомату V и сверхслову β следующим алгоритмом:

1) Изначально $\mathfrak{B}$ пусто, и $u = 1$, u – начальная позиция в слове β, $q = q_0,\ q$ – начальное состояние алгоритма.

2) Возьмем u', $u \le u'$ такое, что $\varphi(q, \beta(u) \dots \beta(u')) = q_{0,i}$ для некоторого i, и при этом для любого u'', $u \le u'' < u'$ выполнено, что $\varphi(q, \beta(u) \dots \beta(u'')) \ne q_{0,i}$ для любого i.

3) Пусть $\varphi(q, \beta(u) \dots \beta(u'))$ было равно $q_{0,h}$. Возьмем u''', $u' < u'''$ такое максимальное натуральное число, что

$$\varphi(q_{0,h}, \beta(u'+1) \dots \beta(u''')) = q_{0,h}, \tag{1}$$

и при этом выполняется

$$\forall\, u'' :\ u' < u'' \le u''' \ \Rightarrow\ \varphi(q_{0,h}, \beta(u'+1) \dots \beta(u'')) \in C_h. \tag{2}$$

Если такого u''' не существует возможны два случая:

а) Условия (1) и (2) выполняются бесконечное число раз, то есть максимума не существует. В этом случае β – периодическое с предпериодом, и тогда возьмем $u''' = u' + 2^{u'} \cdot |C_h|$. Данный выбор гарантирует, что условия (1) и (2) будут выполнены.

б) Условия (1) и (2) никогда не выполняются, тогда просто пропустим этот шаг.

4) Добавим слово $\beta(u) \dots \beta(u')$ в $\mathfrak{B}$, затем, если u''' существовал, добавим слово $\beta(u'+1) \dots \beta(u''')$. После добавления обновим значения u и q: если u''' существовал, то $u = u''' + 1$, $q = \varphi(q, \beta(u) \dots \beta(u'''))$, если не существовал, то $u = u' + 1$, $q = \varphi(q, \beta(u) \dots \beta(u'))$.

5) Вернемся к шагу 2.

Так как степень угадывания сверхслова β равна 1, то автомат заходит во множество $\{C_0, \dots, C_{p-1}\}$ бесконечное число раз, следовательно, алгоритм корректен. Также заметим, что по построению сверхслово β равно последовательной конкатенации элементов $\mathfrak{B}$, отсюда получаем, что $\mathfrak{B}$ – представление β. И еще отметим, что множество $\mathfrak{C}_H(\mathfrak{B})$ состоит из элементов вида $\beta(u'+1) \dots \beta(u''')$.

Возьмем накрытие $(\mathfrak{B}, H)$ и покажем, что для него верно условие теоремы.

Рассмотрим два случая, первый – слово β является периодическим с предпериодом. Заметим, что в этом случае с некоторого момента времени автомат не выходит из размеченного цикла, а в построении $\mathfrak{B}$ число u''' выбиралось таким, что

$$\varliminf_{n\to\infty} \frac{L(\mathfrak{C}_H(\mathfrak{B}), n)}{n} = 1; \ \varliminf_{n\to\infty} \frac{C(\mathfrak{C}_H(\mathfrak{B}), n)}{n} = 0.$$

Откуда сразу следует, что

$$\varliminf_{n\to\infty} \frac{L(\mathfrak{C}_H(\mathfrak{B}),n) - M(H)\cdot C(\mathfrak{C}_H(\mathfrak{B}),n)}{n} = 1.$$

Пусть теперь β не периодическое с предпериодом. Представим каждый момент времени n как сумму $|T_1(n)|$ и $|T_2(n)|$, где $T_1(n)$ – множество тактов, которые соответствуют элементам построения $\mathfrak{B}$ из пункта 3 до момента n, а $T_2(n)$ – множество всех остальных тактов, то есть тех, которые соответствуют элементам из пункта 2 до момента n. Также обозначим $N_1(n)$ – количество угадываний в такты $T_1(n)$, а $N_2(n)$ – в такты $T_2(n)$, получим, что $N_n = N_1(n) + N_2(n)$.

Заметим, что в такты $T_2(n)$ не может быть больше чем $m-1$ угадывания на каждые m тактов, так как после каждого элемента из пункта 3 представления $\mathfrak{B}$ происходит выход из размеченного цикла, что гарантирует хотя бы одно неугадывание, и, в случае не периодического сверхслова, все проходы по размеченным циклам попадают в такты $T_1(n)$. Следовательно, $N_2(n) \le \frac{m-1}{m} \cdot |T_2(n)|$, и $N_1(n) = |T_1(n)|$, откуда получим

$$\sigma_\beta(V) = 1 = \varliminf_{n\to\infty} \frac{N_1(n) + N_2(n)}{n} \le \varliminf_{n\to\infty} \frac{|T_1(n)| + \frac{m-1}{m}\cdot|T_2(n)|}{n} \le 1,$$

так как $|T_1(n)| + |T_2(n)| = n$, из неравенства следует, что

$$\lim_{n\to\infty} \frac{|T_2(n)|}{n} = 0.$$

По определению верно, что $C(\mathfrak{C}_H(\mathfrak{B}),n) \le |T_2(n)|$, а $M(H)$ – фиксированная константа. Получаем

$$\lim_{n\to\infty} \frac{C(\mathfrak{C}_H(\mathfrak{B}),n)}{n} = \lim_{n\to\infty} \frac{M(H)\cdot C(\mathfrak{C}_H(\mathfrak{B}),n)}{n} = 0.$$

Теперь заметим, что $|T_1(n)| \le L(\mathfrak{C}_H(\mathfrak{B}),n) \le n$, следовательно,

$$\sigma_\beta(V) \le \varliminf_{n\to\infty} \frac{L(\mathfrak{C}_H(\mathfrak{B}),n)}{n} = 1.$$

Из двух предыдущих выводов следует, что

$$\varliminf_{n\to\infty} \frac{L(\mathfrak{C}_H(\mathfrak{B}),n) - M(H)\cdot C(\mathfrak{C}_H(\mathfrak{B}),n)}{n} = 1.$$

Необходимость доказана.

Достаточность. Построим автомат $V = (E_k, Q, E_k, \varphi, \psi, q)$, который будет почти полностью прогнозировать β.

Возьмем множество $H = \{\alpha_0, \ldots, \alpha_{p-1}\}$, по построению 1, получим размеченные циклы для каждого слова из H, обозначим их как частично заданные автоматы $C_{\alpha_i} = (E_k, Q_{\alpha_i}, E_k, \varphi_{\alpha_i}, \psi_{\alpha_i}, q_{0,\alpha_i})$, $i \in \{0, \ldots, p-1\}$, $Q_{\alpha_i} = \{q_{0,\alpha_i}, \ldots, q_{|\alpha_i|-1,\alpha_i}\}$. Также из леммы 2 возьмем автомат $A = (E_k, Q_A, \varphi_A, q_0)$, разделяющий вход по H, и обозначим его состояния $Q_A = \{q_0, \ldots, q_{m-1}\}$ и пусть $\{q_{m-p}, \ldots, q_{m-1}\}$ – состояния, сопоставленные словам из H, причем слову α_i сопоставлено состояние q_{m-p+i} для любого i.

Далее приступим к построению автомата V:

1) Положим $Q = Q_{\alpha_0} \cup \cdots \cup Q_{\alpha_{p-1}} \cup Q_A \setminus \{q_{m-p}, \ldots, q_{m-1}\}$, за начальное состояние q примем состояние q_0.

2) Определим функцию переходов, рассмотрим несколько случаев:

   а) Пусть $q' \in Q_A \setminus \{q_{m-p}, \ldots, q_{m-1}\}$ и $\varphi_A(q', a) = q_i$. Если $i \in \{0, \ldots, m-p-1\}$, то положим $\varphi(q', a) = q_i$. Если же $i \in \{m-p, \ldots, m-1\}$, то $\varphi(q', a) = q_{0,\alpha_i}$.

   б) Пусть $q' \in Q_{\alpha_i}$ для некоторого $i \in \{0, \ldots, p-1\}$. Если для символа a, $a \in E_k$, функция $\varphi_{\alpha_i}(q', a)$ определена, то положим $\varphi(q', a) = \varphi_{\alpha_i}(q', a)$.

   в) Теперь рассмотрим случай, когда $q' \in Q_{\alpha_i}$ для некоторого $i \in \{0, \ldots, p-1\}$, и функция $\varphi_{\alpha_i}(q', a)$ не определена. Пусть $q' = q_{j,\alpha_i}$, $j \in \{0, \ldots, |\alpha_i|-1\}$. Рассмотрим значение $\varphi_A(q_0, \alpha_i\,]_j\, a) = q_{i'}$. Если $i' \in \{0, \ldots, m-p-1\}$, то положим $\varphi(q', a) = q_{i'}$. Если $i' \in \{m-p, \ldots, m-1\}$, то положим $\varphi(q', a) = q_{0,\alpha_{i'}}$.

3) Функцию выхода зададим разметкой. В состояниях $Q_{\alpha_0} \cup \cdots \cup Q_{\alpha_{p-1}}$ отметим то же ребро, что и в соответствующем цикле, в остальных состояниях отметим произвольное ребро.

Далее рассмотрим степень прогнозирования автоматом V сверхслова β. Как только в сверхслове β встречается подслово, являющееся элементом из $\mathfrak{C}_H(\mathfrak{B})$, то не более чем через $M(H)$ тактов автомат начинает угадывать и угадывает на протяжении всей длины элемента $\mathfrak{B}$. Это обеспечивается пунктом 2в построения автомата и тем, что автомат A приходит в состояние, сопоставленное некоторому слову из H, не более чем через $M(H)$ тактов после подачи соответствующего слова. Смена размеченных циклов происходит тоже не более чем за $M(H)$ тактов по пункту 2в и построению автомата A. Отсюда получаем, что $N_n \geq L(\mathfrak{C}_H(\mathfrak{B}), n) - M(H) \cdot C(\mathfrak{C}_H(\mathfrak{B}), n)$, и, следовательно,

$$\sigma_\beta(V) \geq \varliminf_{n\to\infty} \frac{L(\mathfrak{C}_H(\mathfrak{B}), n) - M(H) \cdot C(\mathfrak{C}_H(\mathfrak{B}), n)}{n} = 1.$$

Теорема доказана.

□

Список литературы

[1] Вереникин А.Г., Гасанов Э.Э., "Об автоматной детерминизации множеств сверхслов", *Дискретная математика*, **18**:2 (2006), 84–97.

[2] Гасанов Э.Э., "Прогнозирование периодических сверхсобытий автоматами", *Интеллектуальные системы*, **19**:1 (2015), 23–34.

[3] Мастихина А.А., "О частичном угадывании сверхслов", *Интеллектуальные системы*, **11**:1–4 (2007), 561–572.

[4] Мастихина А.А., "Критерий частичного предвосхищения общерегулярных свехсобытий", *Дискретная математика*, **23**:4 (2011), 103–114.

[5] Ведерников И.К., "Исследование алгоритма, задающего функцию выхода прогнозирующего автомата", *Интеллектуальные системы*, **20**:3 (2016), 103–111.

[6] Кудрявцев В.Б., Алешин С.В., Подколзин А.С., *Введение в теорию автоматов*, «Наука», Москва, 1985, 320 с.

[7] Alfred V. Aho, Margaret J. Corasick, "Efficient string matching: An aid to bibliographic search", *Communications of the ACM*, **18**:6 (1975), 333–340

Criterion for almost complete prediction of a superword in a multi-valued alphabet

Vedernikov I.K.

The machine predicts the next character of the input sequence if it outputs that character the moment before.

The present paper considers whether an arbitrary superword in multivalued alphabet can be almost completely predicted or not.

The paper provides a theorem that enables to restrict the class of machines, with the help of which superwords are predicted. Moreover, the paper presents the criterion for almost complete predicting.

Keywords: almost complete predicting, predicting machine, prediction of superwords by a machine, criterion for predicting.

References

[1] Verenikin A.G., Gasanov E.E., "On automata determination of sets of superwords", *Discrete Math*, **18**:2 (2006), 84–97

[2] Gasanov E.E., "Prediction of periodic super-events by automata", *Intelligent systems*, **19**:1 (2015), 23–34

[3] Mastihina A.A., "About Partial Guessing of Superwords", *Intelligent systems*, **11**:1–4 (2007), 561–572

[4] Mastihina A.A., “The criterion for partial anticipation of general regular superevents”, *Discrete Math*, **23**:4 (2011), 103–114

[5] Vedernikov I.K., “Investigation of the Algorithm Specifying the Output Function of the Predicting Automaton”, *Intelligent systems*, **20**:3 (2016), 103–111

[6] Kudryavtsev V.B., Alyoshin S.V., Podkolzin A.S., *Introduction to automata theory*, «Science», Moscow, 1985, 320 c.

[7] Alfred V. Aho, Margaret J. Corasick, “Efficient string matching: An aid to bibliographic search”, *Communications of the ACM*, **18**:6 (1975), 333–340