Предугадывание сверхслов на отрезке

Маншилин О.Г.[1]

Автомат предугадывает символ входного сверхслова, если он выдаёт этот символ на выходе в предыдущий момент времени.

В работе вводится понятие степени предугадывания на отрезке. Исследуется вопрос о взаимосвязи предугадывания на отрезке и бесконечности. Получен результаты, позволяющие судить о степени предугадывания на отрезке по степени предугадывания на бесконечности и о степени предугадывания на бесконечности по степени предугадывания на отрезке.

Ключевые слова: предугадывающий автомат, автоматное предугадывание общерегулярных сверхслов, степень предугадывания на отрезке.

1. Введение

В этой статье рассматривается предвосхищение общерегулярных сверхсобытий над алфавитом $\{0, 1\}$. Ещё на заре развития теории автоматов было введено понятие предугадывания. Позже было показано, что идеально предугадываются только полностью периодические сверхслова [1]. Это ограничение можно обойти используя более сложную структуру автоматов [3].

Подобную задачу рассматривали не только с точки зрения конечных автоматов, но также с точки зрения теории вероятностей и машинного обучения [4]. Однако в нашей работе все сверхслова предугадываются конечными автоматами.

В работе Мастихиной А.А. было введено понятие частичного предугадывания сверхслов на бесконечности [2]. В этой статье введено понятие

[1] *Маншилин Олег Григорьевич* — магистр каф. высшей математики ф-та фундаментальных наук МГТУ им. Н.Э.Баумана, e-mail: manshilin.o@gmail.com.

Manshilin Oleg Grigorievich— graduate student, Bauman Moscow State University, Faculty of Fundamental Science, Chair of Higher Math.

частичного предугадывания сверхслов на отрезке. Автомат верно предугадывает некоторую долю входных символов на отрезке длиной l, которая называется степенью предугадывания на отрезке и определяется как нижний нижняя грань отношения угаданных символов на отрезке длиной l к l по всем возможным отрезкам длиной l в сверхслове.

Была дана оценка степени предугадывания на отрезке для конечных автоматов определённого вида. Была показана связь между предвосхищением сверхслов на отрезке и предвосхищением сверхслов на бесконечности для конечных автоматов произвольного вида.

В этой статье приведены примеры конечных автоматов, на которых явно видна связь между предвосхищением на бесконечности и предвосхищением на отрезке. Часть теорем, приведённых в этой статье, распространяется как на общерегулярные сверхслова так и на контекстно-свободные сверхслова.

Автор выражает благодарность А.А. Мастихиной за постановку задачи и помощь в работе.

2. Основные понятия и формулировка результатов.

В работе используются нижеприведённые понятия:

- $c_{\infty}^{\mathfrak{A}}$ - степень предугадывания автомата $\mathfrak{A}$ на бесконечности.
- $|a|$ - длина слова $a \in A^*$.
- $\alpha(n)$ - n-ый элемент слова или сверхслова α
- $pref(\alpha, n)$ - префикс слова или сверхслова α: $pref(\alpha, n) = \alpha(1)\alpha(2)...\alpha(n)$
- $a \cdot b$ - конкатенация слов a и b
- a^n - повторения слова a n раз.
- $\lfloor x \rfloor$ - округления действительного числа x до целочисленного в нижнюю сторону.

В работе рассматриваются конечные инициальные автоматы:

$$\mathfrak{A} = (A, Q, B, \varphi, \psi, q_0),$$

где A - входной алфавит, Q - множество состояний автомата, B - выходной алфавит, φ - функция переходов, ψ - функция выходов.

Если на вход автомату A подаётся сверхслово α, на выходе получается сверхслово y, а состояние автомата в момент времени t обозначается как q_t, то функционирование автомата задаётся системой:

$$\begin{cases} y(t) = \psi(\alpha(t), q_{t-1}) \\ q_t = \varphi(\alpha(t), q_{t-1}) \end{cases}$$

Выходное сверхслово автомата $\mathfrak{A}$ при подаче на него сверхслова α будем обозначать как $y_\alpha^{\mathfrak{A}}$. На протяжение всей статьи множество $\{0, 1\}$ рассматривается как входной и выходной алфавиты.

Мы говорим, что автомат $\mathfrak{A}$ частично угадывает сверхслово на бесконечности [2], если:

$$|y_{\alpha(i)}^{\mathfrak{A}} - \alpha(i+1)| < \infty \tag{1}$$

Степень предугадывания на бесконечности определяется как [2]:

$$c_\infty^\theta(\alpha) = 1 - \underline{\lim}_{t\to\infty} \frac{\sum_1^t |y_\alpha^\theta(i) - \alpha(i+1)|}{t} \tag{2}$$

Будем говорить, что автомат частично предугадывает сверхслово α в смысле отрезка, когда выполняется неравенство:

$$\sum_{i=N+n}^{N+n+t_1} |y_\alpha^{\mathfrak{A}}(i) - \alpha(i+1)| < t_1 | \text{для } \forall n \in \mathbb{N}, \tag{3}$$

где N - конечный префикс сверхслова α, t_1 - длина отрезка, на котором предугадывается сверхслово α. Это означает, что по прошествии некого конечного префикса на любом подслове длиной t_1 будет угадан по меньшей мере один символ.

Введём степень предугадывания в смысле отрезка как:

$$c_{t_1}^\theta(\alpha) = \sup_{N\in\mathbb{N}} \inf_{n\in\mathbb{N}} \{1 - \frac{\sum_{N+n}^{N+n+t_1} |y_\alpha^\theta(i) - \alpha(i+1)|}{t_1}\} \tag{4}$$

Множество сверхслов называется частично предугадываемым на бесконечности, если существует такой конечный автомат, на котором степень предугадывания каждого сверхслова множества строго больше нуля.

Множество сверхслов R называется частично предугадываемым на отрезке длиной t_1, если существует такой конечный автомат, для которого степень предугадывания на любом произвольном подслове длиной t_1 на каждом сверхслове множества R строго больше нуля.

Автомат $\mathfrak{A}$ предугадывает произвольное множество сверхслов $S \subset \{0,1\}^\infty$ на бесконечности со степенью θ, если для $\forall \alpha, \alpha \in S, c_\infty^{\mathfrak{A}} \geqslant \theta$.

Автомат $\mathfrak{A}$ предугадывает произвольное множество сверхслов $S \subset \{0,1\}^\infty$ на отрезке длиной t_1 со степенью θ, если для $\forall \alpha, \alpha \in S, c_{t_1}^{\mathfrak{A}} \geqslant \theta$

Регулярное событие над алфавитом A определяется как:

- $\emptyset, \{a\}, a \in A$ - регулярные события.
- Пусть R_1, R_2 - регулярные события. Тогда $R_1 \cdot R_2, R_1 \cup R_2$ и R_1^* тоже являются регулярными. R_1^* является множеством всех возможных конечных итераций события R_1.

Общерегулярное событие над алфавитом A определяется как:

- Если R_1 - регулярное событие над алфавитом A, то R_1^∞ - общерегулярное событие над этим же алфавитом. R_1^∞ является сверхитерацией события R_1.
- Если R_1 - регулярное событие над алфавитом A, а R_2 - общерегулярное событие над алфавитом A, то $R_1 \cdot R_2$ - общерегулярное событие над алфавитом A.
- Если R_1, R_2 -общерегулярные сверхсобытия над алфавитом A, то $R_1 \cup R_2$ - общерегулярное сверхсобытие над алфавитом A.

Конечный автомат $\mathfrak{A}$ принимает множество сверхслов R_1^∞ с помощью множеств состояний $M = \{M_1, \ldots, M_t\}$, если автомат проходит все состояния из одного $M_i \in M$ бесконечное число раз и проходит оставшиеся состояния конечное число раз.

В любом автомате можно выделить сильно связные компоненты $Q_1, ..., Q_k$, то есть совокупность состояний, достижимых друг из друга $\forall q^{'}, q^{''} \in Q_i, i = 1, ..., k, \exists \alpha \in \{0,1\}^*$, т.ч. $\varphi(q^{'}, \alpha) = q^{''}$.

Назовём сильно связную компоненту замкнутой, если из неё не достижима никакая другая сильно связная компонента. Очевидно, что хотя бы одна замкнутая сильно связная компонента существует.

Тупиковым множеством будем называть такую сильно связную компоненту $Q^{'}$, что $\forall Q^{''} \in 2^{Q^{'}}, Q^{''} \not\subseteq M$.

Будем обозначать множество состояний, из которых самое близкое тупиковое множество [2] достигается за l рёбер как $Q(l)$

Теорема 1. *Имеется конечный автомат θ, принимающий множество сверхслов R. Пусть у этого автомата присутствуют тупиковые множества и максимальная длина пути до тупикового множества равна n_1. Подобному автомату можно задать выходную функцию таким образом, что его степень предугадывания множества R на отрезке длиной l будет больше или равна $\frac{\lfloor l/n_1 \rfloor}{l}$.*

Теорема 2. *Степень предугадывания конечного автомата на бесконечности всегда больше степени предугадывания конечного автомата на отрезке.*

Теорема 3. *Если у конечного автомата $\mathfrak{A}$ степень предугадывания на бесконечности строго больше нуля, то найдётся такая длина отрезка l, что степень предугадывания автомата $\mathfrak{A}$ на отрезке длиной l будет строго больше нуля.*

3. Вывод теорем

3.1. Вывод степени предугадывания на отрезке для предугадывающего автомата, полученного из принимающего

Теорема 1. *Имеется конечный автомат θ, принимающий множество сверхслов R. Пусть у этого автомата присутствуют тупиковые множества и максимальная длина пути до тупикового множества равна n_1. Подобному автомату можно задать выходную функцию таким образом, что его степень предугадывания множества R на отрезке длиной l будет больше или равна $\frac{\lfloor l/n_1 \rfloor}{l}$.*

Доказательство. Зададим функцию выхода автомату θ следующим образом:

- **1-ый шаг:**

  Для всех состояний из множества $Q(1)$ определим выходные функции следующим образом: если для состояния $q \in Q(1)$ и входного символа $x \in \{0,1\}$ справедливо что $\varphi(q,x) \in T$, то $f(q) = \overline{x}$.

- **i-ый шаг:**

  Для всех состояний из множества $Q(i)$ определим выходные функции следующим образом: если для состояния $q \in Q(i)$ и входного символа $x \in \{0,1\}$ справедливо что $\varphi(q,x) \in Q(i-1)$, то $f(q) = \overline{x}$

Для получения степени предугадывания на отрезке построим наихудшую последовательность символов для предугадывания. Ради этой цели начнём строить последовательность из состояния $q_0 \in Q(n_1)$. В состоянии q_0 мы можем либо угадать следующий символ и остаться во множестве состояний $Q(n_1)$, либо не угадать и перейти в состояние $Q(n_1 - 1)$. Верное предугадывание гарантируется лишь во множестве состояний $Q(1)$. Действительно, если из каждого множества $Q(i)$ автомат будет переходить во множество $Q(i-1)$, то он перейдёт во множество $Q(1)$ за n_1 шагов из состояния q_0. Наихудшим переходом из множества $Q(1)$ является переход во множество $Q(n_1)$. При переходе во множество $Q(n_1)$ автомат может ошибиться $n_1 - 1$ раз подряд в наихудшем случае, когда как при переходе во множество $Q(i), i < n_1$ автомат может ошибиться $i - 1$ раз подряд в наихудшем случае. Это означает, что при переходе во множество $Q(i), i < n_1$, автомат будет реже ошибаться чем при переходе во множество $Q(n_1)$, а значит степень предугадывания при таком построении не будет являться минимальной. При прохождении n_1 символов обязательно угадывается лишь один символ из них. Это соответствует тому, что при прохождении l символов гарантированно будет угадано лишь $\lfloor l/n_1 \rfloor$ символов. Теорема доказана. □

На Рис. 1 показан пример конечного автомата, который предугадывает множество сверхслов $(0(11 \cup 10)^*0)^\infty$ со степенью 0.5 на бесконечности, но с меньшей степенью на отрезке любой длины.

Рассмотрим слово $(00)^\infty$. Степень предугадывания на бесконечности этого сверхслова автоматом, изображённым на рис. 1, равняется 0.5. Степень предугадывания такого слова на отрезках длиной кратной двум также равняется 0.5.

Теперь рассмотрим слово $(010000)^\infty$. Степень предугадывания этого сверхслова автоматом, показанным на рис. 1, равняется 0.5 на бесконечности. Степень предугадывания данного автомата на отрезке длиной в 3 символа равняется $\frac{1}{3}$. Степень предугадывания данного автомата на отрезке длиной в 4 символа равняется $\frac{1}{4}$.

На Рис. 2 показан пример конечного автомата, который предугадывает множество сверхслов $(01^*00^*1)^\infty$ со степенью $0.(3)$ на бесконечности и на отрезках длиной кратной трём.

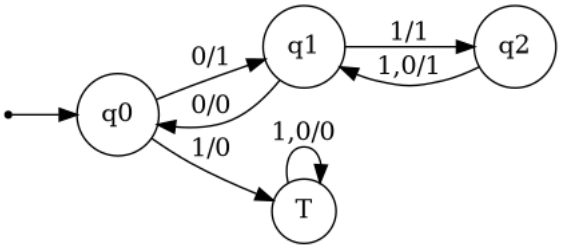

Рис. 1. Пример конечного автомата, у которого степени предугадывания на отрезке и бесконечности не совпадают.

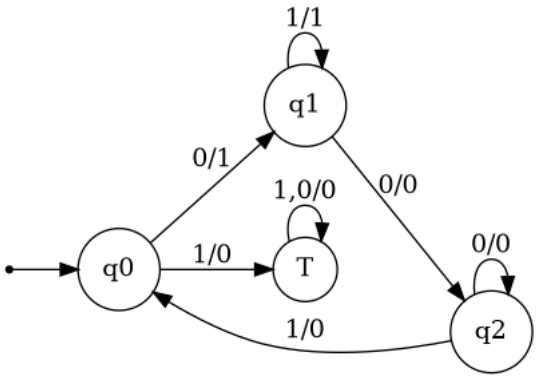

Рис. 2. Пример конечного автомата, у которого степени предугадывания на отрезке и бесконечности совпадают.

3.2. Вывод степени предугадывания на бесконечности из степени предугадывания на отрезке

Теорема 2. *Степень предугадывания конечного автомата на бесконечности всегда больше степени предугадывания конечного автомата на отрезке.*

Доказательство. Выпишем определение степени предугадывания на отрезке.

$$c^{\theta}_{t_1}(\alpha) = \sup_{N\in\mathbb{N}} \inf_{n\in\mathbb{N}} \{1 - \frac{\sum_{N+n}^{N+n+t_1} |y^{\theta}_{\alpha}(i) - \alpha(i+1)|}{t_1}\},$$

где N - длина конечного префикса, $t_1 \in \mathbb{N}$ - длина отрезка, на котором производится предугадывание.

Раскроем infinum.

$$S_{t_1} = \sum_{1}^{t1} |y_\alpha^{\mathfrak{A}}(i) - \alpha(i+1)| \leqslant t_1 - t_1 c_{t_1}^{\mathfrak{A}}(\alpha) \tag{5}$$

Выпишем определение степени предугадывания на бесконечности и раскроем предел.

$$c_\infty^{\mathfrak{A}}(\alpha) = 1 - \underline{\lim}_{t\to\infty} \frac{\sum_1^t |y_\alpha^{\mathfrak{A}}(i) - \alpha(i+1)|}{t}$$

$$c_\infty^{\mathfrak{A}}(\alpha) = 1 - \underline{\lim}_{t\to\infty} \frac{\sum_1^N |y_\alpha^{\mathfrak{A}}(i) - \alpha(i+1)| + \sum_{N+1}^{t/t_1} S_{t_1}}{t} =$$
$$1 - \underline{\lim}_{t\to\infty} \frac{(\frac{t}{t1} - N - 1)S_{t_1}}{t} = 1 - \frac{S_{t_1}}{t_1}$$

Заменим S_{t_1} неравенством (5).

$$c_\infty^{\mathfrak{A}}(\alpha) = 1 - \frac{S_{t_1}}{t_1} \geqslant 1 - \frac{t_1 - t_1 c_{t_1}^{\mathfrak{A}}(\alpha)}{t_1} = 1 - 1 + c_{t_1}^{\theta}(\alpha) = c_{t_1}^{\mathfrak{A}}(\alpha)$$

$$c_\infty^{\mathfrak{A}} \geqslant c_{t_1}^{\mathfrak{A}}$$

Теорема доказана. □

Замечание: В доказательстве не использовалось никаких уникальных свойств конечных автоматов, вследствие чего **теорема 2** распространяется как на конечные автоматы, так и на произвольные детерминированные автоматы.

На Рис. 3 приведён пример конечного автомата, который угадывает контекстно-свободное сверхслово $(0^n1^n)^\infty$ со степенью $\frac{1}{2}$ как на бесконечности, так и на отрезке, длина которого строго больше или равна двум.

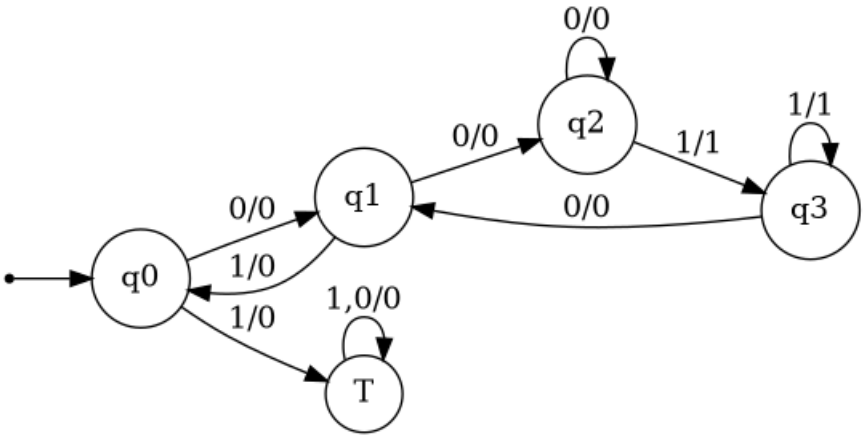

Рис. 3. Пример конечного автомата, который предугадывает контекстно свободный язык как на бесконечности, так и на отрезке.

3.3. Вывод степени предугадывания на отрезке из степени предугадывания на бесконечности для произвольного конечного автомата

Теорема 3. *Если у конечного автомата $\mathfrak{A}$ степень предугадывания на бесконечности строго больше нуля, то найдётся такая длина отрезка l, что степень предугадывания автомата $\mathfrak{A}$ на отрезке длиной l будет строго больше нуля.*

Доказательство. Покажем, что для любого конечного автомата $\mathfrak{A}$, который предугадывает множество сверхслов R_1 с нулевой степенью на отрезке любой длины, можно построить сверхслово из множества R_1, для которого $c_\infty^{\mathfrak{A}} = 0$. Возьмём конечный автомата $\mathfrak{A}$, предугадывающий множество сверхслов R_1. Допустим, что его степень предугадывания равна нулю для любой конечной длины отрезка: $c_l^{\mathfrak{A}} = 0$. Вместе с предугадывающим конечным автоматом $\mathfrak{A}$ рассмотрим принимающий конечный автомат $\mathfrak{B}$. Подавая одни и те же символы на вход конечных автоматов $\mathfrak{A}$ и $\mathfrak{B}$, сопоставим между собой пути в автомате $\mathfrak{A}$ и пути в автомате $\mathfrak{B}$. Все циклы автомата $\mathfrak{B}$ будем брать соответствующими одному конкретному сверхслову, принадлежащему R_1, степень предугадывания которого равна нулю на любой длине отрезка.

В начальный момент автомат $\mathfrak{A}$ будет находится в состоянии $q_{k_0}^{\mathfrak{A}}$, а автомат $\mathfrak{B}$ - в состоянии $q_{k_0}^{\mathfrak{B}}$. Будем подавать на оба автомата такую по-

следовательность символов α, на которой конечный автомат $\mathfrak{A}$ не предугадывает ни единого символа. Обозначим длину α как l: $|\alpha| = l$. Такую последовательность можно найти любой длины благодаря предположению, что степень предугадывания автомата $\mathfrak{A}$ на отрезке любой длины равна нулю. Получим последовательности состояний $\theta_1^{\mathfrak{A}} = q_{k_0}^{\mathfrak{A}} q_{k_1}^{\mathfrak{A}} q_{k_2}^{\mathfrak{A}} ... q_{k_l}^{\mathfrak{A}}$ для автомата $\mathfrak{A}$ и последовательность состояний $\theta_1^{\mathfrak{B}} = q_{k_0}^{\mathfrak{B}} q_{k_1}^{\mathfrak{B}} q_{k_2}^{\mathfrak{B}} ... q_{k_l}^{\mathfrak{B}}$ для автомата $\mathfrak{B}$, где $q_{k_i} = \varphi(pref(\theta_1, i), \alpha(i))$. Различные q_{k_i} могут обозначать одно и то же состояние. Подберём такую длину последовательности α и такое число $n \in \mathbb{N}, n < l$, что как $\theta_2^{\mathfrak{A}} = q_{k_n}^{\mathfrak{A}} q_{k_{n+1}}^{\mathfrak{A}} ... q_{k_l}^{\mathfrak{A}}$ так и $\theta_2^{\mathfrak{B}} = q_{k_n}^{\mathfrak{B}} q_{k_{n+1}}^{\mathfrak{B}} ... q_{k_l}^{\mathfrak{B}}$ будут являться циклами. Будем обозначать как α_n такую подпоследовательность слова α, при подаче которой на автоматы $\mathfrak{A}$ и $\mathfrak{B}$ они проходят циклы $\theta_2^{\mathfrak{A}}$ и $\theta_2^{\mathfrak{B}}$ соответственно. Такое n и слово α всегда возможно подобрать ввиду того, что слово α с нулевой степенью предугадывания можно найти произвольной длины, ввиду нулевой степени предугадывания на отрезке любой длины. Цикл же в любом конечном автомате всегда можно получить подав на него достаточно большое количество символов. В наихудшем случае он последовательно пройдёт все свои состояния, а затем перейдёт в одно из ранее пройденных состояний, чем образует цикл. Следующим шагом будем повторять циклы $\theta_2^{\mathfrak{A}}$ и $\theta_2^{\mathfrak{B}}$. При таком построение возможно два варианта:

- Сверхслово $pref(\alpha, n)(\alpha_n)^\infty$ проходит все состояния, которыми конечный автомат принимает множество сверхслов R_1. При подаче такого сверхслова на предугадывающий конечный автомат $\mathfrak{A}$ он будет бесконечно проходить цикл $\theta_2^{\mathfrak{A}}$, на котором он по построению не предугадывает ни одного символа. При этом сверхслово $pref(\alpha, n)(\alpha_n)^\infty$ принадлежит множеству R^∞. Вследствие этого степень предугадывания сверхслов $pref(\alpha, n)(\alpha_n)^\infty$ на бесконечности равняется нулю: $c_\infty^{\mathfrak{A}} = 0$.

- Сверхслово $pref(\alpha, n)(\alpha_n)^\infty$ не проходит все состояния, которыми конечный автомат $\mathfrak{B}$ принимает множество R_1. Тогда будем строить требуемое сверхслово таким образом:

  $$pref(\alpha, n)\alpha_n\beta\alpha_n\alpha_n\beta...(\alpha_n)^m\beta...,$$

  где β является словом по которому как принимающий так и предугадывающий автоматы проходят цикл. Слово β подбирается таким образом, чтобы при его подаче автомат $\mathfrak{B}$ прошёл все оставшиеся

состояния, необходимые для принятия сверхслова, а также конечное количество произвольных принимающих состояний. Такое слово β всегда можно подобрать ввиду конечного количества состояний у предугадывающего и принимающего автоматов. Сверхслово $pref(\alpha, n)\alpha_n\beta\alpha_n\alpha_n\beta...(\alpha_n)^m\beta...$ по построению принадлежит множеству R_1. Его степень предугадывания на бесконечности равняется нулю, вследствие стремления m к бесконечности и абсолютной непредугадываемости цикла $\theta_2^{\mathfrak{A}}$.

Раз можно построить по меньшей мере одно свехслово из множества со степенью предугадывания на бесконечности, равной нулю, то степень предугадывания на бесконечности у всего множества равняется нулю. Теорема доказана. □

Утверждение 1. *Для контекстно-свободных языков из предугадывания на бесконечности не следует предугадывание на отрезке.*

Приведём пример: Возьмём конечный автомат, выдающий на любой входной символ константу ноль. Попробуем предугадать им сверхслово $(0^n1^n)^\infty$. Очевидно, что степень предугадывания такого сверхслова этим автоматом будет равняться 0.5. Однако подслово 1^n может иметь любую конечную длину и более того, оно повторяется в сверхслове. Из-за повторений подслова его невозможно пропустить, выбросив конечное число символов из начала сверхслова. Получается, что насколько большой отрезок мы не брали бы - всегда можно найти такую последовательность 1^n, что её длина будет больше нежели длина отрезка, на котором мы предугадываем сверхслово. Подобный отрезок может повторяется бесконечное число раз в сверхслове. Из этого следует, что степень предугадывания на отрезке у такого автомата равняется нулю.

Итого мы получили пару автомат - сверхслово, у которой степень предугадывания на бесконечности равняется 0.5, а степень предугадывания на отрезке равняется нулю.. Данное наблюдение доказывает утверждение.

Список литературы

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

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

[3] Tim Smith, “Prediction of Infinite Words with Automata”, 2016, arXiv: https://arxiv.org/abs/1603.02597.

[4] S. E. Marzen, J. P. Crutchfield, "Probabilistic Deterministic Finite Automata and Recurrent Networks, Revisited", 2019, arXiv: https://arxiv.org/abs/1910.07663.

Prediction of superword segments

Manshilin O.G.

The automaton predicts the character of the input sequence, if it outputs this character at the previous time.

The paper introduces the concept of the degree of prediction on a superword segments. The question of the relationship between prediction on superword segments and prediction on superwords is investigated.

We obtained results that allow us to judge the degree of prediction on superword segments if we know degree of prediction on superword and vice versa.

Keywords: predicting automaton, prediction of superwords with automatons, prediction degree on finite subsequency.