О частичном угадывании сверхслов

А. А. Мастихина

В работе исследуется угадывающий автомат, который в каждый момент времени на выходе старается предугадать значение входа в следующий момент. Доля угаданных букв во входном слове называется степенью угадывания автомата. В работе для различных степеней угадывания получены условия существования угадываемых и неугадываемых сверхслов для двух случаев: когда они являются периодическими и непериодическими.

1. Введение

Понятие угадывающего автомата впервые было введено в статье [1].

Автомат угадывает сверхслово a над алфавитом $\{0, 1\}$, если при подаче a на вход автомат с некоторого момента времени начинает на своем выходе предугадывать значение входного слова в следующий момент.

В данной статье рассматривается частичное угадывание, для этого введено понятие степени угадывания сверхслова автоматом, как доля верно предсказанных входных символов. Были получены следующие результаты.

Для полного угадывания периодического слова достаточно автомата с числом состояний, равным длине периода.

Чтобы угадать периодическое слово меньше чем наполовину, достаточно автомата с одним состоянием.

Для любого натурального n можно найти такое сверхслово, что ни один автомат с n состояниями не сможет угадать его более чем наполовину.

Построено сверхслово, которое ни один автомат не может угадать ни с какой степенью, а также сверхслово, которое любой автомат угадывает наполовину, и доказано, что такого сверхслова, которое каждый автомат угадывает больше чем наполовину, не существует.

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

2. Основные понятия и формулировка результатов

Будем использовать следующие обозначения:

- $\{0,1\}^n$ — множество всех слов длины n в алфавите $\{0,1\}$,
- $\{0,1\}^*$ — множество всех слов в алфавите $\{0,1\}$, по определению будем считать, что пустое слово Λ принадлежит $\{0,1\}^*$,
- $\{0,1\}^\infty$ — множество всех сверхслов в алфавите $\{0,1\}$,
- $|a|$ — длина слова $a \in \{0,1\}^*$. По определению $|\Lambda| = 0$,
- $a(n)$ — n-ый элемент слова или сверхслова a,
- $a\,\underset{n}{]}$ — префикс a длины n, то есть $a\,\underset{n}{]} = a(1)\ldots a(n)$,
- ab — конкатенация слов a и b,
- $a^n = \underbrace{a\ldots a}_{n}$, здесь n — натуральное и может быть равно ∞,
- $A(n)$ — множество периодических сверхслов из алфавита $\{0,1\}$ с периодом, равным n и конечным предпериодом, то есть $A(n) = \{ap^\infty : p \in \{0,1\}^n, a \in \{0,1\}^*\}$.

В статье рассматриваются конечные инициальные автоматы следующего вида:

$$\mathfrak{A} = (\{0,1\}, Q, \{0,1\}, \varphi, \psi, q_0)\,,$$

где $\{0,1\}$ — входной алфавит, Q — множество состояний, которое является конечным подмножеством некоторого фиксированного счётного множества, $\{0,1\}$ — выходной алфавит, $\varphi : \{0,1\} \times Q \to Q$ — функция переходов, $\psi : 0,1 \times Q \to \{0,1\}$ — функция выходов, q_0 — начальное состояние.

Если на вход автомату $\mathfrak{A}$ подается сверхслово x, на выходе получается сверхслово y, и q_t означает состояние автомата в момент времени t, то функционирование автомата задается системой

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

Далее выходное сверхслово автомата $\mathfrak{A}$ при подаче на его вход сверхслова a будем обозначать через $y_a^{\mathfrak{A}}$.

Автомат $\mathfrak{A}$ *угадывает сверхслово* $a \in \{0,1\}^\infty$, если

$$\sum_{i=1}^{\infty} |y_a^{\mathfrak{A}}(i) - a(i+1)| < \infty.$$

Если $a \in \{0,1\}^\infty$, то обозначим

$$d^{\mathfrak{A}}(a,t) = \frac{1}{t}\sum_{i=1}^{t}(1 - |y_a^{\mathfrak{A}}(i) - a(i+1)|),$$

$$c^{\mathfrak{A}}(a) = \varliminf_{t\to\infty} d^{\mathfrak{A}}(a,t).$$

Автомат $\mathfrak{A}$ *угадывает сверхслово* a *со степенью* α, если

$$c^{\mathfrak{A}}(a) \geqslant \alpha.$$

Пусть $\mathfrak{B}_\alpha(a)$ — множество всех автоматов, угадывающих слово a со степенью α.

Обозначим через $\omega(B)$ число состояний автомата B и введем

$$R_\alpha(a) = \min_{B\in\mathfrak{B}_\alpha} \omega(B)$$

— минимальное число состояний автомата, угадывающего слово a со степенью α.

Пусть случайная величина ξ, с одинаковой вероятностью принимает значения 0 или 1. Тогда каждое сверхслово $s = s(1)s(2)\ldots s(i)\ldots$ такое, что $s(i)$ для любого $i \in \mathbb{N}$ есть реализация случайной величины ξ, назовем *статистическим.*

Теорема 1.

1) *Если $\alpha = 1$, то для любого $n \in \mathbb{N}$ и любого $b \in A(n)$ выполнено*

$$\frac{m}{2} \leqslant R_\alpha(b) \leqslant m \leqslant n,$$

где m — длина минимального периода сверхслова b.

2) *Если* $\alpha \leqslant \dfrac{1}{2}$, *то для любого* $n \in \mathbb{N}$ *и любого* $b \in A(n)$ *выполнено*

$$R_\alpha(b) = 1.$$

3) *Если* s *— некоторое статистическое сверхслово,* $\alpha > \dfrac{1}{2}$, *то для любого* $n \in \mathbb{N}$ *и любого* $\delta > 0$ *найдутся такие числа* $n_1, n_2 \in \mathbb{N}$, *что для сверхслова* $b = s(1)\ldots s(n_1)(s(n_1+1)\ldots s(n_2))^\infty$ *с вероятностью не менее* $1-\delta$ *выполнено* $R_\alpha(b) > n$.

Теорема 2. *Для любого* $\alpha \in (0,1]$ *существует такое сверхслово, что ни один автомат не угадывает его со степенью* α.

Теорема 3.

1) *Для любого* $\alpha \in \left[0, \frac{1}{2}\right]$ *с вероятностью* 1 *любое статистическое сверхслово угадывается со степенью* α.
2) *Ни для какого* $\alpha \in \left(\frac{1}{2}, 1\right]$ *не существует такого сверхслова, что все автоматы угадывают его со степенью* α.

3. Угадывание периодических сверхслов

Доказательство теоремы 1

1) Покажем, что для любого $n \in \mathbb{N}$ и любого $b \in A(n)$ существует автомат не более, чем с n состояниями, угадывающий b со степенью $\alpha = 1$. Рассмотрим периодическую часть сверхслова: $p(1)\ldots p(m), m \leqslant n$. Построим автомат с состояниями $q_1, \ldots, q_m$, и его функции ψ и φ не зависят от входной последовательности, а устроены так:

$$\begin{cases} \varphi(x, q_i) = q_{i+1},\ i = 1...m-1, \\ \varphi(x, q_m) = q_1, \\ \psi(x, q_i) = p(i+1),\ i = 1...m-1, \\ \psi(x, q_m) = p(1). \end{cases}$$

Определение начального состояния автомата зависит от длины предпериода сверхслова. Если длина предпериода равна $km + j$, то в качестве начального состояния берется $q_0 := q_{j-1}$. Такой автомат удовлетворяет требованиям теоремы.

Нижняя оценка следует из [1, теорема 2].

2) Рассмотрим два автомата с одним состоянием: выдающий константу 0 или константу 1. Обозначим их соответственно $\mathfrak{A}_0$ и $\mathfrak{A}_1$. Пусть угадываемое слово имеет вид $b = ap^\infty$. Если в слове p больше нулей, то автомат $\mathfrak{A}_0$ будет угадывающим со степенью не менее $\dfrac{1}{2}$, в противном случае угадывающим будет — $\mathfrak{A}_1$.

3) Рассмотрим произвольный автомат $B = (\{0,1\}, \{0,1\}, \phi, \psi)$. Его функции выхода никак не зависят от следующего символа последовательности и наоборот. Следовательно, вероятность их совпадения, если слово статистическое, равна $\dfrac{1}{2}$.

Тогда число $s_t^{\mathfrak{B}} = |y_s^{\mathfrak{B}}(t) - s(t+1)|$ для каждого $t \in \mathbb{N}$ также можно рассматривать как реализацию случайной величины ξ, значит, их последовательность $s_1^{\mathfrak{B}} s_2^{\mathfrak{B}} \ldots s_t^{\mathfrak{B}} \ldots$ является статистическим сверхсловом.

Возьмем произвольное натуральное n.

Перечислим все неэквивалентные автоматы с числом состояний, не большим n, которых конечное число M: $\mathfrak{B}_1, \mathfrak{B}_2, \ldots, \mathfrak{B}_M$.

Введем $\mathfrak{q}$ — вектор состояний всех автоматов $\mathfrak{B}_1, \mathfrak{B}_2, \ldots, \mathfrak{B}_M$ на статистическом сверхслове s.

При подаче t символов s вектор состояний определяется так:

$$\mathfrak{q}_t = (q_t^{\mathfrak{B}_1}, \ldots, q_t^{\mathfrak{B}_M}).$$

Так как таких векторов конечное число (наборы конечного числа состояний), на бесконечном слове некоторые из них будут повторятся бесконечное число раз. Возьмем один из таких векторов и обозначим его $\mathfrak{q}^*$, а момент времени, когда $\mathfrak{q}^*$ встречается впервые, обозначим n_1.

Каждое статистическое сверхслово есть последовательность реализаций одной случайны величины ξ, математическое ожидание которой равно $M\xi = \frac{1}{2}$. По закону больших чисел

$$\frac{1}{t}\sum_{i=1}^{t} s_i \to \frac{1}{2},$$

при $t \to \infty$ по вероятности, то есть для любого сколь угодно малого $\varepsilon > 0$

$$\lim_{t\to\infty} P\{|\frac{1}{t}\sum_{i=1}^{t} s_i - \frac{1}{2}| \geqslant \varepsilon\} = 0.$$

Из определения $s_j^{\mathfrak{B}_i}$ следует, что $d^{\mathfrak{B}_i}(s,t) = \frac{1}{t}\sum_{j=1}^{t} s_j^{\mathfrak{B}_i}$, $i = 1,\dots,M$.

Так как к математическому ожиданию сходится бесконечная сумма, конечную ее часть можно отбросить. Введем

$$\overline{d}^{\mathfrak{B}_i}(s,t) = \frac{1}{(t-n_1)} \sum_{j=n_1+1}^{t} s^{\mathfrak{B}_i}.$$

Ясно, что $\overline{d}^{\mathfrak{B}_i}(s,t)$ для любого сколь угодно малого $\varepsilon > 0$ удовлетворяет

$$\lim_{t\to\infty} P\{|\overline{d}^{\mathfrak{B}_i}(s,t) - \frac{1}{2}| \geqslant \varepsilon\} = 0.$$

Тогда для любых сколь угодно малых $\delta > 0$ и $\varepsilon > 0$ для каждого автомата $\mathfrak{B}_i$, такого, что $w(\mathfrak{B}_i) \leqslant n$ найдется такой момент времени $t_0^i = t_0^i(\varepsilon,\delta)$, что для любых $t \geqslant t_0^i$ выполнено

$$P\{|\overline{d}^{\mathfrak{B}_i}(s,t) - \frac{1}{2}| < \varepsilon\} \leqslant \delta.$$

Выберем ε так, чтобы $\alpha > \frac{1}{2} + \varepsilon$.

Теперь возьмем $t_0(\varepsilon,\delta) = \max\{t_0^1(\varepsilon,\delta), t_0^2(\varepsilon,\delta), \dots, t_0^M(\varepsilon,\delta)\}$. Пусть n_2 — первый момент, больший чем $\max(t_0(\varepsilon,\delta), n_1)$, в который вектор состояний $\mathfrak{q}$ снова снова принимает значение $\mathfrak{q}^*$. На отрезке от n_1 до n_2 все рассмотренные автоматы угадывают примерно половину символов (а именно меньше, чем $\frac{1}{2}+\varepsilon$). Таким образом с n_1 до n_2 все M автоматов совершат цикл, в котором все они угадывали $\frac{1}{2}+\varepsilon < \alpha$ символов.

Тогда с вероятностью $1-\delta$ сверхслово $b = s(1)\dots s(n_1)(s(n_1+1) \dots s(n_2))^\infty$ каждый автомат с числом состояний, не большим n, будет угадывать со степенью меньшей α. Значит, для угадывания со степенью α состояний нужно больше.

Третье утверждение теоремы 1 доказано.

4. Угадывание непериодических сверхслов

Теперь будем рассматривать непериодические сверхслова $a \in \{0,1\}^\infty$.

Доказательство теоремы 2

Рассмотрим произвольный инициальный автомат $\mathfrak{B} = (\{0,1\}, Q, \{0,1\}, \varphi, \psi, q_0)$.

Расширим функции φ и ψ на $\{0,1\}^* \times Q$, а именно, если на вход автомату $\mathfrak{B}$ подается слово x, то $\varphi(x\underset{t}{]}, q_0)$, $\psi(x\underset{t}{]}, q_0)$ означают состояние, в котором находится автомат, и выходное значение в момент времени t, соответственно.

Возьмем произвольное слово s конечной неотрицательной длины. Если подать его на вход автомата $\mathfrak{B}$, то мы получим другой инициальный автомат $\mathfrak{B}' = (\{0,1\}, Q, \{0,1\}, \varphi, \psi, \varphi(s, q_0))$. Через $N(\mathfrak{B}, s)$ будем обозначать сверхслово полностью неугадываемое автоматом $\mathfrak{B}'$.

Формально сверхслово $N(\mathfrak{B}, s)$ определяется индуктивно следующим образом.

$$N(\mathfrak{B}, s)(1) = \begin{cases} \overline{\psi}(s, q_0), & \text{если } |s| > 0, \\ 0, & \text{если } |s| = 0. \end{cases}$$

При $t > 1$

$$N(\mathfrak{B}, s)(t) = \overline{\psi}(N(\mathfrak{B}, s)\underset{t-1}{]}, \varphi(s, q_0)).$$

Возьмем все инициальные конечные автоматы (их счетное число) и занумеруем: $\mathfrak{A}_1, \mathfrak{A}_2, \ldots, \mathfrak{A}_k, \ldots$

Для $t \in \mathbb{N}$ через R_t обозначим слово над алфавитом $\mathbb{N}$, состоящее из первых t натуральных чисел, то есть $R_t = 1, 2, \ldots, t$.

Построим сверхслово r, полученное последовательной конкатенацией слов $R_1 R_2 \ldots R_t \ldots$ То есть r имеет вид:

$$12123123412345\ldots 12\ldots t\ldots$$

и схематично изображено на рисунке 1.

Возьмем произвольное $\alpha > 0$. Теперь для заданного α построим сверхслово l, которое не будет угадываться со степенью α ни одним автоматом.

Возьмем натуральную константу $C > \frac{1}{\alpha}$ и определим l как конкатенацию слов $L_1, L_2, L_3, \ldots, L_i, \ldots$, где

$$L_i = N(\mathfrak{A}_{r(i)}, L_1 L_2 L_3 \ldots L_{i-1})\underset{C^i - C^{i-1}}{]}.$$

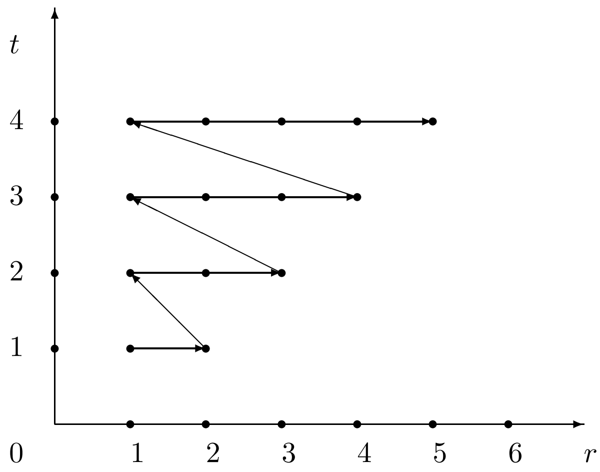

Рис. 1. Сверхслово r.

То есть сверхслово l есть последовательность отрезков полностью неугадываемых сверхслов для соответствующих автоматов (порядок которых задается сверхсловом r), причем каждый отрезок длиннее предыдущего в C раз.

Длина префикса $L_1 \ldots L_i$ будет равна

$$C + (C^2 - C) + (C^3 - C^2) + \ldots + (C^i - C^{i-1}) = C^i.$$

Введем вспомогательную последовательность j_t^m такую, что для любого $t \in \mathbb{N}$ выполнено $r(j_t^m) = m$.

В сверхслове r число m первый раз встречается, когда проходит

$$\underbrace{1, 2,}_{2} \underbrace{1, 2, 3,}_{3} \ldots \underbrace{1, 2, \ldots, m}_{m},$$

то есть $2 + 3 + 4 + \ldots + m$ символов. Значит,

$$j_1^m = \frac{m(m+1)}{2} - 1.$$

Далее m встречается через m шагов, потом через $m + 1$; расстояние между каждыми следующими символами m будет на единицу больше. Следовательно, при $t > 1$

$$j_t^m = \frac{m(m+1)}{2} - 1 + m(t-1) + \sum_{k=0}^{t-2} k =$$
$$= \frac{m(m+1)}{2} - 1 + m(t-1) + \frac{(t-2)(t-1)}{2}.$$

Для каждого автомата $\mathfrak{A}_m$, $m \in \mathbb{N}$, рассмотрим последовательность n_t^m такую, чтобы n_t^m-тый символ сверхслова l был последней буквой в $L_{j_t^m}$ для любого $t \in \mathbb{N}$, то есть

$$n_t^m = C^{j_t^m} = C^{\frac{m(m+1)}{2} - 1 + m(t-1) + \frac{(t-2)(t-1)}{2}}.$$

Рассмотрим число угаданных символов автоматом $\mathfrak{A}_m$ на подпоследовательности n_t^m. Если предположить, что весь отрезок $L_1 L_2 \ldots L_{j_t^m - 1}$ угадан, то, учитывая, что $L_{j_t^m}$ по построению полностью этим автоматом не угадывается, получим

$$c^{\mathfrak{A}_m}(l) \leqslant \lim_{t\to\infty} \frac{1}{n_t^m} \cdot \sum_{i=1}^{n_t^m} (1 - |y_l^{\mathfrak{A}_m}(i) - l(i+1)|) <$$
$$< \lim_{t\to\infty} \frac{|L_1 \ldots L_{j_t^m - 1}|}{|L_1 \ldots L_{j_t^m}|} = \lim_{t\to\infty} \frac{C^{j_t^m - 1}}{C^{j_t^m}} = \frac{1}{C} < \alpha.$$

Теорема 2 доказана.

Видно, что существование неугадываемого сверхслова основано на нижнем пределе в определении степени угадывания. Если бы степень угадывания определялась как нижний предел $d^{\mathfrak{A}}(a, t)$, можно было бы для любой степени угадывания построить сверхслово, напротив, угадываемое всеми автоматами.

В данном же случае сверхслово, угадываемое всеми автоматами, существует не всегда.

Доказательство теоремы 3

1) $\alpha \leqslant \dfrac{1}{2}$.

Рассмотрим произвольное статическое сверхслово $s = s(1)s(2)\ldots s(t)\ldots$ Так как его буквы есть независимые реализации случайной величины ξ, любая подпоследовательность символов статистического слова также является статистическим.

Следовательно, для любого автомата $\mathfrak{A}$ на любой подпоследовательности n_t $\forall \varepsilon > 0$ выполняется

$$\lim_{t\to\infty} P\{|d^{\mathfrak{A}}(s, n_t) - \frac{1}{2}| \geqslant \varepsilon\} = 0.$$

То есть с вероятностью 1 любой автомат угадывает статистическое сверхслово со степенью $\frac{1}{2}$.

2) Теперь рассмотрим случай $\frac{1}{2} < \alpha \leqslant 1$. Для него достаточно рассмотреть 2 автомата: выдающий константу 0 и константу 1. Обозначим их соответственно $\mathfrak{A}_0$ и $\mathfrak{A}_1$.

Предположим существует сверхслово a, которое оба эти автомата угадывают со степенью α. Рассмотрим последовательность n_t^0, на которой достигается нижний предел $d^{\mathfrak{A}_0}(a, t)$.

Тогда для автомата $\mathfrak{A}_1$ имеем

$$c^{\mathfrak{A}_1}(a) \leqslant \lim_{t\to\infty} d^{\mathfrak{A}_1}(a, n_t^0) = 1 - \lim_{t\to\infty} d^{\mathfrak{A}_0}(a, n_t^0) = 1 - \alpha < \alpha.$$

Полученное противоречие доказывает теорему 3.

Список литературы

[1] Вереникин А. Г., Гасанов Э. Э. Об автоматной детерминизации множеств сверхслов // Дискретная математика. 2006. Т. 18. № 2. С. 84–97.

[2] Кудрявцев В. Б., Алешин С. В., Подколзин А. С. Введение в теорию автоматов. М.: Наука, 1985.

[3] Ширяев А. Н. Вероятность. М.: Наука, 1980.