1 Основные определения

Стохастической КС-грамматикой [3] называется система $G =< V_T, V_N, R, s >$, где V_T и V_N — конечные множества терминальных и нетерминальных символов (терминалов и нетерминалов) соответственно, $s \in V_N$ — аксиома, R — множество правил. Множество R можно представить в виде $R = \cup_{i=1}^{n} R_i$, где n — мощность алфавита V_N и $R_i = \{r_{i1}, \dots, r_{in_i}\}$. Каждое правило r_{ij} из R_i имеет вид

$$r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij}, \qquad j = 1, \dots, n_i, \tag{1}$$

где $A_i \in V_N$, $\beta_{ij} \in (V_N \cup V_T)^*$ и p_{ij} — вероятность применения правила r_{ij}, причём

$$0 < p_{ij} \le 1, \qquad \sum_{j=1}^{n_i} p_{ij} = 1. \tag{2}$$

Для $\alpha, \gamma \in (V_N \cup V_T)^*$ будем говорить, что γ выводится из α (и обозначать $\alpha \Rightarrow \gamma$), если существуют $\alpha_1, \alpha_2 \in (V_N \cup V_T)^*$, для которых $\alpha = \alpha_1 A_i \alpha_2$, $\gamma = \alpha_1 \beta_{ij} \alpha_2$ и в грамматике имеется правило $A_i \xrightarrow{p_{ij}} \beta_{ij}$. Через $\Rightarrow_*$ обозначим рефлексивное транзитивное замыкание отношения $\Rightarrow$. Грамматика G задаёт контекстно-свободный язык $L_G = \{\alpha \in V_T^* : s \Rightarrow_* \alpha\}$. Будем говорить, что слово α *выводимо* грамматикой G, если $\alpha \in L_G$.

Выводом слова α назовём последовательность правил $\omega(\alpha) = (r_{i_1 j_1}, r_{i_2 j_2}, \dots, r_{i_q j_q})$, с помощью последовательного применения которых слово α выводится из аксиомы s. Если при этом каждое правило применяется к самому левому нетерминалу в слове, такой вывод называется левым. Для вывода $\omega(\alpha) = (r_{i_1 j_1}, \dots, r_{i_q j_q})$ определим величину $p(\omega(\alpha)) = p_{i_1 j_1} \cdot \ldots \cdot p_{i_q j_q}$.

Каждое слово, выводимое грамматикой G, имеет *дерево вывода* [4]. Дерево вывода для слова α строится следующим образом. Корень дерева помечается аксиомой s. Далее последовательно рассматриваются правила левого вывода слова α. Пусть на очередном шаге рассматривается правило $A_i \xrightarrow{p_{ij}} b_{i_1} b_{i_2} \dots b_{i_m}$, где $b_{i_l} \in (V_N \cup V_T)$ $(l = 1, \dots, m)$. Тогда из самой левой вершины-листа дерева, помеченной символом A_i, проводится m дуг в вершины следующего яруса, которые помечаются слева направо символами $b_{i1}, \dots, b_{i,m}$ соответственно. После построения дуг и вершин для всех правил в выводе листья дерева помечены терминальными символами (либо пустым словом λ, если применяется правило вида $A_i \xrightarrow{p_{ij}} \lambda$) и само слово получается при обходе листьев дерева слева направо. *Высотой* дерева вывода будем называть максимальную длину пути от корня к листу.

Пример

Рассмотрим пример КС-грамматики G, задающей язык арифметических выражений $+, *$ без скобок с параметрами a и b.

$$\begin{array}{lcl} G & = & < V_N, V_T, S, R > \\ V_N & = & \{S, T, M\} \\ V_T & = & \{+, *, a, b\} \end{array} \tag{3}$$

Множество R правил вывода содержит правила:

$$\begin{array}{lllll} r_{11} & : & S & \to & T \\ r_{12} & : & S & \to & T+S \\ r_{21} & : & T & \to & M \\ r_{22} & : & T & \to & M*T \\ r_{31} & : & M & \to & a \\ r_{32} & : & M & \to & b \end{array} \tag{4}$$

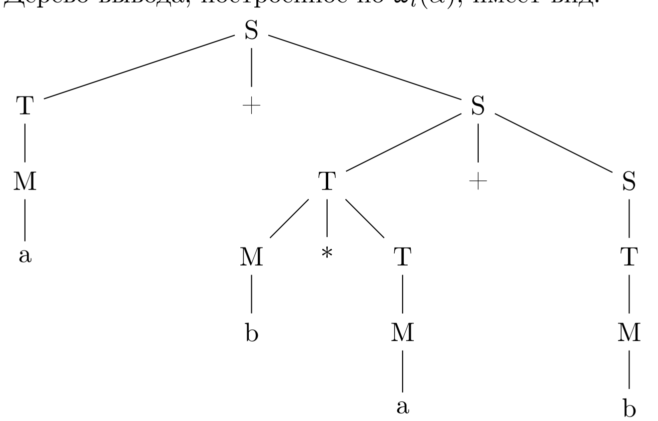
Рассмотрим слово $\alpha = a+b*a+b$, выводимое грамматикой G. Левый вывод этого слова имеет вид:

$$\omega_l(\alpha) = (r_{12}, r_{21}, r_{32}, \ldots) \tag{5}$$

Последовательно применяя правила левого вывода к аксиоме S грамматики, получим слово α:

$$\begin{aligned} S \to T+S \to M+S \to a+S \to a+T+S \to a+M*T+S \to \\ \to a+b*T+S \to a+b*M+S \to a+b*a+S \to \\ \to a+b*a+T \to a+b*a+M \to a+b*a+b \end{aligned} \tag{6}$$

Дерево вывода, построенное по $\omega_l(\alpha)$, имеет вид:



Обозначим $p(\alpha) = \sum p(\omega_l(\alpha))$, где сумма берётся по всем левым выводам слова α. Грамматика G называется *согласованной*, если

$$\lim_{n\to\infty} \sum_{\substack{\alpha\in L_G \\ |\alpha|\le n}} p(\alpha) = 1. \tag{7}$$

Согласованная грамматика G задаёт распределение вероятностей P на множестве L_G, при этом $p(\alpha)$ — вероятность слова α. Пара $\mathcal{L} = (L_G, P)$ называется *стохастическим КС-языком.* В дальнейшем будем всюду предполагать, что рассматривается согласованная грамматика.

Будем говорить, что нетерминал A_j *непосредственно выводится* из нетерминала A_i, и обозначать $A_i \to A_j$, если в грамматике имеется правило $A_i \xrightarrow{p_{ij}} \alpha_1 A_j \alpha_2$, где $\alpha_1, \alpha_2 \in (V_N \cup V_T)^*$. Рефлексивное транзитивное замыкание отношения $\to$ обозначим $\to_*$. Будем говорь, что нетерминал A_j *выводится* из A_i, если $A_i \to_* A_j$. Если одновременно $A_i \to_* A_j$ и $A_j \to_* A_i$, будем обозначать $A_i \leftrightarrow_* A_j$. Отношение эквивалентности $\leftrightarrow_*$ разбивает множество нетерминалов грамматики на классы

$$K_1, K_2, \ldots, K_m. \tag{8}$$

Множества номеров нетерминалов, входящих в класс K_j обозначим через I_j. Грамматика называется *разложимой* при $m \ge 2$, и *неразложимой* в противном случае.

Будем говорить, что класс K_j *непосредственно следует* за классом K_i, и обозначать $K_i \prec K_j$, если $i \ne j$ и существуют такие $A_1 \in K_i$ и $A_2 \in K_j$, что $A_1 \to A_2$. Рефлексивное транзитивное замыкание отношения $\prec$ обозначим $\prec_*$, и будем говорить, что класс K_j *следует* за классом K_i, если $K_i \prec_* K_j$. Отношение $\prec_*$ задаёт частичный порядок на множестве классов $K_1, \ldots, K_m$.

Назовём класс K *особым*, если он содержит ровно один нетерминал A_i, и в грамматике отсутствует правило вида $A_i \xrightarrow{p_{ij}} \alpha_1 A_i \alpha_2$, где $\alpha_1, \alpha_2 \in (V_N \cup V_T)^*$. В дальнейшем всюду будем предполагать, что грамматика не содержит особых классов.

2 Производящие функции

Пусть $\alpha \in (V_N \cup V_T)^*$ — слово в объединённом алфавите терминальных и нетерминальных символов. Через $l_i(\alpha)$ будем обозначать число нетерминалов A_i в слове α, а через $l(\alpha)$ — характеристический вектор $(l_1(\alpha), l_2(\alpha), \ldots, l_k(\alpha))$.

Введём вероятностные производящие функции $F_i(\mathbf{s})$:

$$F_i(\mathbf{s}) = F_i(s_1, s_2, \ldots, s_k) = \sum_{j=1}^{n_i} p_{ij} s_1^{l_1} s_2^{l_2} \cdot \ldots \cdot s_k^{l_k}, \tag{9}$$

где суммирование происходит по всем правилам вывода r_{ij} из R_i, и $l_s = l_s(\beta_{ij})$ — число нетерминалов A_s в правой части β_{ij} правила r_{ij}.

Производящие функции $F_i(\mathbf{s})$ содержат информацию о том, с какой вероятностью мы можем получить слово с тем или иным характеристическим вектором в результате однократного применения случайного правила r_{ij} к нетерминалу A_i. При этом правило выбирается в соответствии с распределением вероятностей p_{ij}. Если в $F_i(\mathbf{s})$ присутствует слагаемое вида $ps_1^{l_1} \ldots s_k^{l_k}$, значит слово с характеристическим вектором $l = (l_1, \ldots, l_k)$ будет получено с вероятностью p.

Для удобства будем обозначать $\mathbf{F}(\mathbf{s}) = (F_1(\mathbf{s}), \ldots, F_k(\mathbf{s}))$.

Введём производящие функции $F_i(t, \mathbf{s})$ с параметром t:

$$F_i(t, \mathbf{s}) = F_i(t, s_1, s_2, \ldots, s_k) = \begin{cases} F_i(t-1, F(\mathbf{s})), & \text{при } t > 1 \\ F_i(\mathbf{s}), & \text{при } t = 1 \end{cases} \tag{10}$$

Производящие функции $F_i(t, \mathbf{s})$ содержат информацию о том, с какой вероятностью мы можем получить слово с определённым характеристическим вектором $l = (l_1, l_2, \ldots, l_k)$ в результате построения t ярусов дерева вывода с корнем в нетерминале A_i.

---

3 Моменты. Матрица первых моментов грамматики

Величины

$$a_j^i = \left. \frac{\partial F_i(\mathbf{s})}{\partial s_j} \right|_{\mathbf{s}=\mathbf{1}} \tag{11}$$

называются *первыми моментами*, и определяют математическое ожидание числа нетерминалов A_j в слове, полученном в результате однократного применения случайного правила вывода к нетерминалу A_i.

Аналогично введём величины $a^i_j(t)$:

$$a^i_j(t) = \left.\frac{\partial F_i(t,\mathbf{s})}{\partial s_j}\right|_{\mathbf{s}=\mathbf{1}} \tag{12}$$

Величины $a^i_j(t)$ определяют математическое ожидание числа нетерминалов A_j в слове, полученном в результате построения t ярусов дерева вывода из нетерминала A_i.

Мы будем также рассматривать вторые b^i_{jl} и третьи c^i_{jln} моменты

$$b^i_{jl} = \frac{\partial^2 F_i(\mathbf{s})}{\partial s_l \partial s_j}, \qquad c^i_{jln} = \frac{\partial^3 F_i(\mathbf{s})}{\partial s_n \partial s_l \partial s_j}, \tag{13}$$

а также величины $b^i_{jl}(t)$, $c^i_{jln}(t)$:

$$b^i_{jl}(t) = \frac{\partial^2 F_i(t,\mathbf{s})}{\partial s_l \partial s_j}, \qquad c^i_{jln}(t) = \frac{\partial^3 F_i(t,\mathbf{s})}{\partial s_n \partial s_l \partial s_j}, \tag{14}$$

Матрица A, составленная из элементов a^i_j, называется *матрицей первых моментов* грамматики.

4 Вероятности продолжения

Список литературы

[1] **Шеннон К.** Математическая теория связи. М.: ИЛ, 1963

[2] **Марков А. А.** Введение в теорию кодирования. М.: Наука, 1982

[3] **Фу К.** Структурные методы в распознавании образов. М.: Мир, 1977

[4] **Ахо А., Ульман Дж.** Теория синтаксического анализа, перевода и компиляции. Том 1. М.: Мир, 1978

[5] **Севастьянов Б. А.** Ветвящиеся процессы. — М.: Наука, 1971 — 436 с.

[6] **Гантмахер Ф. Р.** Теория матриц. — 5-е изд., — М.: ФИЗМАТЛИТ, 2010

[7] **Жильцова Л. П.** О матрице первых моментов разложимой стохастической КС-грамматики. УЧЁНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА, Том 151, кн. 2, 2009

[8] **Жильцова Л. П.** Закономерности применения правил грамматики в выводах слов стохастического контекстно-свободного языка // Математические вопросы кибернетики. Выр. 9. М.: Наука, 2000. С. 100-126.

[9] **Жильцова Л. П.** О нижней оценке стоимости кодирования и асимптотически оптимальном кодировании стохастического контекстно-свободного языка // Дискретный анализ и исследование операций. Серия 1, т. 8, №3. Новосибирск: Издательство Института математики СО РАН, 2001. С. 26-45.

[10] **Борисов А. Е.** Закономерности в словах стохастических контекстно-свободных языков, порождённых грамматиками с двумя классами нетерминальных символов. Вопросы экономного кодирования. // Диссертация на соискание учёной степени кандидата физико-математических наук. Нижний Новгород, 2006.