О свойствах деревьев вывода для стохастической КС-грамматики, имеющей вид «цепочки»

Л.П. Жильцова, И.М. Мартынов (Нижний Новгород)

В работе исследуются свойства деревьев вывода высоты t при $t \to \infty$ для стохастической КС-грамматики с разложимой матрицей A первых моментов специального вида. Рассматривается критический случай, когда перронов корень матрицы A равен 1.

Стохастической КС–грамматикой называется система $G = \langle V_T, V_N, R, s \rangle$, где V_T и V_N — конечные алфавиты терминальных и нетерминальных символов (терминалов и нетерминалов) соответственно, $s \in V_N$ — аксиома, $R = \cup_{i=1}^{k} R_i$, где k — мощность алфавита V_N и R_i — множество правил с одинаковой левой частью A_i. Каждое правило r_{ij} из R_i имеет вид

$$r_{ij} : A_i \overset{p_{ij}}{\to} \beta_{ij},\ j = 1, ..., n_i,$$

где $A_i \in V_N$, $\beta_{ij} \in (V_T \cup V_N)^*$ и p_{ij} – вероятность применения правила r_{ij}, причем $0 < p_{ij} \leqslant 1$ и $\sum_{j=1}^{n_i} p_{ij} = 1$.

Применение правила грамматики к слову состоит в замене вхождения нетерминала из левой части правила на слово, стоящее в его правой части.

Каждому слову α КС-языка соответствует последовательность правил грамматики (вывод), с помощью которой α выводится из аксиомы s. Выводу слова соответствует дерево вывода [1], вероятность которого определяется как произведение вероятностей правил, образующих вывод.

По стохастической КС-грамматике строится матрица A первых моментов. Для нее элемент a_j^i определяется как $\sum_{l=1}^{n_i} p_{il} s_{il}^j$, где величина s_{il}^j равна числу нетерминальных символов A_j в правой части правила r_{il}. Перронов корень матрицы A обозначим через r.

Введем некоторые обозначения. Будем говорить, что нетерминал A_j непосредственно следует за нетерминалом A_i (и обозначать $A_i \to A_j$), если в грамматике существует правило вида $A_i \overset{p_{ij}}{\to} \alpha_1 A_j \alpha_2$, где $\alpha_1, \alpha_2 \in (V_T \cup V_N)^*$. Рефлексивное транзитивное замыкание отношения $\to$ обозначим $\to_*$.

Классом нетерминалов назовем максимальное по включению подмножество $K \subseteq V_N$ такое, что $A_i \to_* A_j$ для любых $A_i, A_j \in K$. Для различных классов нетерминалов K_1 и K_2 будем говорить, что класс K_2 непосредственно следует за классом K_1 (и обозначать $K_1 \prec K_2$), если существуют $A_1 \in K_1$ и $A_2 \in K_2$, такие, что $A_1 \rightarrow A_2$. Рефлексивное транзитивное замыкание отношения $\prec$ обозначим через $\prec_*$.

Пусть $\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$ — множество классов нетерминалов грамматики, $m \geqslant 2$. Будем полагать, что классы нетерминалов перенумерованы таким образом, что $K_i \prec_* K_j$ тогда и только тогда, когда $i < j$.

Будем говорить, что грамматика имеет вид «цепочки», если ее матрица первых моментов A имеет вид

$$A = \begin{pmatrix} A_{11} & A_{12} & 0 & \cdots & 0 & 0 \\ 0 & A_{22} & A_{23} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_{m-1,m-1} & A_{m-1,m} \\ 0 & 0 & 0 & \cdots & 0 & A_{m,m} \end{pmatrix}. \tag{1}$$

Один класс нетерминалов представлен в матрице множеством подряд идущих строк и соответствующим множеством столбцов с теми же номерами. Для класса K_i квадратная подматрица, образованная соответствующими строками и столбцами, обозначается через A_{ii}. Подматрица A_{ij} является нулевой, если $K_i \not\prec K_j$. Блоки, расположенные ниже главной диагонали, нулевые в силу упорядоченности классов.

Для грамматики с матрицей первых моментов вида (1) классы нетерминалов образуют линейный порядок по отношению $\prec$:

$$K_1 \prec K_2 \prec \ldots \prec K_i \prec \ldots \prec K_m. \tag{2}$$

Для каждого класса K_i матрица A_{ii} неразложима. Без ограничения общности будем считать, что она строго положительна и непериодична. Обозначим через r_i перронов корень матрицы A_{ii}. Для неразложимой матрицы перронов корень является вещественным и простым [2]. Очевидно, $r = \max_i\{r_i\}$.

Пусть $J = \{i_1, i_2, \ldots, i_l\}$ — множество всех номеров i_j классов, для которых $r_{i_j} = 1$. Рассмотрим подцепочку классов

$$K_j \prec K_{j+1} \prec \ldots \prec K_m. \tag{3}$$

Число классов с номерами из J в такой цепочке обозначим через q_j.

Через $P_j(t)$ обозначим вероятность множества деревьев вывода высоты t, корень которых помечен нетерминалом A_j.

Теорема 1 *При $t \to \infty$*

$$P_j(t) \sim U^{(j)} \cdot \frac{c_j}{t^{1+\left(\frac{1}{2}\right)^{q_j-1}}},$$

где c_j - некоторая положительная константа.

При $r_j = 1$ вектор $U^{(j)}$ является правым собственным вектором для матрицы A_{jj}, соответствующим r_j.

Обозначим через $M_{ij}(t)$ математическое ожидание числа применений правила r_{ij} грамматики в дереве вывода высоты t, корень которого помечен аксиомой грамматики $s = A_1$.

Теорема 2 *Пусть матрица первых моментов грамматики G имеет вид (1), и r_{ij} — правило грамматики, для которого $A_i \in K_l$. Тогда при $t \to \infty$*

$$M_{ij}(t) \sim d_i \cdot p_{ij} \cdot t^{\left(\frac{1}{2}\right)^{q_l^* - 1}},$$

где $q_l^* = q_l - 1$ при $l \in J$ и $q_l^* = q_l$ при $l \notin J$; $d_i > 0$ – некоторая константа, и p_{ij} — вероятность правила r_{ij}.

Таким образом, величина $M_{ij}(t)$ определяется удаленностью класса K_l, которому принадлежит нетерминал A_i из левой части правила r_{ij}, от конца цепочки (2). Математическое ожидание $M_{ij}(t)$ тем больше, чем меньше число классов с номерами из множества J, следующих за классом K_l в (3). Следовательно, чем дальше удален класс от начала цепочки (2), тем чаще применяются соответствующие ему правила грамматики. Для последнего класса в (2) с номером из J и всех последующих классов величины $M_{ij}(t)$ соответствующих правил грамматики имеют порядок $O\left(t^2\right)$, как в случае неразложимой грамматики [3] и грамматики с двумя классами нетерминалов [4].

Список литературы

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

[2] Гантмахер Ф.Р. Теория матриц. — М.: ФИЗМАТЛИТ, 2010. — 560 с.

[3] Жильцова Л.П. Закономерности в деревьях вывода слов стохастического контекстно-свободного языка и нижняя оценка стоимости кодирования. Критический случай// Дискретный анализ и исследование операций. Серия 1, т.10, N3. Новосибирск: Издательство Института математики СО РАН, 2003. С.23-53.

[4] Борисов А.Е. О свойствах слов языка, порожденного разложимой стохастической КС-грамматикой с двумя нетерминалами. Критический случай// Материалы VIII Международного семинара "Дискретная математика и ее приложения". М.: Изд. мех-мат. ф-та МГУ, 2004. С. 408-410.