О распределении нетерминалов в деревьях вывода стохастической КС-грамматики вида «цепочки»

И. М. Мартынов

murbidodrus@gmail.com

ННГУ им. Н. И. Лобачевского, Нижний Новгород

В работе исследуются вероятностные свойства деревьев вывода стохастической КС-грамматики специального вида. Рассматривается случай, когда матрица первых моментов A грамматики разложима и имеет перронов корень равный 1. Целью работы являвляется исследование распределения нетерминальных символов в деревьях вывода высоты t, при $t \to \infty$.

Стохастической КС–грамматикой называется система $G = \langle V_N, V_N, R, s\rangle$, где V_T и V_N — конечные алфавиты терминальных и нетерминальных символов 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 одинкаовой левой частью вида

$$r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij},\ j = 1, 2, \ldots, n,$$

где $A_i \in V_N$, $\beta_{ij} \in (V_T \cup V_N)^*$ и p_{ij} — вероятность применения правила r_{ij}, причём $0 < p_{ij} \leqslant 1$ 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, вероятность которого определяется как произведение вероятностей правил, образующих вывод.

По стохастической КС-грамматике строится матрица A первых моментов. Для неё 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.

Введём отношение на множестве нетерминалов. Будем обозначать $A_i \to A_j$, если в грамматике существует правило вида $A_i \xrightarrow{p_{il}} \alpha_1 A_j \alpha_2$, где $\alpha_1, \alpha_2 \in (V_T \cup V_N)^*$. Рефлексивное транзитивное замыкание отношения $\to$ обозначим $\to_*$.

Классом нетерминалов назовём максимальное по включению подмножество $K \in V_N$ такое, что $A_i \to_* A_j$ для любых $A_i, A_j \in K$. Отношение $\to_*$ на множестве нетерминалов порождает отношение на множестве их классов. Будем обозначать $K_1 \prec K_2$, если существуют $A_1 \in K_1$ и $A_2 \in K_2$, такие, что $A_1 \to A_2$. 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 говорить, что грамматика имеет вид «цепочки», если $K_i \prec K_j$ тогда и только тогда, когда $i + 1 = j$.

Матрица первых моментов грамматики вида «цепочки» имеет вид:

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

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} $(i \neq j)$ является нулевой, если $K_i \not\prec K_j$. Блоки, расположенные ниже главной диагонали, нулевые в силу упорядоченности классов.

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 [1]. Очевидно, $r = \max\limits_i \{r_i\}$.

Пусть $J = \{i_1, i_2, \ldots, i_l\}$ — множество всех номеров i_j классов, для которых $r_{ij} = 1$. Классы K_l такие, что $l \in J$, будем называть критическими. Также обозначим через q_l число критических классов среди подцепочки $K_l, K_{l+1}, \ldots, K_m$. Тогда верна следующая теорема.

Теорема 1. *Математические ожидания $M_i(t)$ числа нетерминалов A_i в деревьях вывода высоты t, порождённых стохастической КС-грамматикой вида «цепочки», при $t \to \infty$ удовлетворяют условию:*

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

где $q_l^* = q_l - 1$ при $l \in J$, и $q_l^* = q_l$ при $l \notin J$, $A_i \in K_l$, и d_i — некоторые константы.

Обозначим через $q_i(t)$ число нетерминалов A_i в случайном дереве высоты t, порождённом грамматикой. Рассмотрим произвольную пару нетерминалов A_i и A_j таких, что математические ожидания $M_i(t)$ и $M_j(t)$ имеют один порядок по t, при $t \to \infty$. Верна следующая теорема.

Теорема 2. *Для любых двух нетерминалов $A_i \in K_l$, $A_j \in K_s$ таких, что $q_l^* = q_s^*$, при $t \to \infty$ выполняется условие:*

$$D\left(\frac{q_i(t)}{q_j(t)} - \frac{d_i}{d_j}\right) \to 0,$$

где $q_i(t)$, $q_j(t)$ — количество нетерминалов A_i, A_j соответственно в деревьях высоты t, d_i и d_j — некоторые константы.

Другими словами, при $t \to \infty$ соотношение частот между двумя нетерминалами, имеющими одинаковую асимптотику математических ожиданий, становится всё ближе к фиксированному значению.

Литература

[1] *Гантмахер Ф. Р.* Теория матриц. 5-е изд. — М.: ФИЗМАТЛИТ, 2010. — 560 с.