

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

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

В работе исследуются вероятностные свойства деревьев вывода высоты t при $t \rightarrow \infty$ для стохастической КС-грамматики специального вида. Рассматривается случай, когда матрица первых моментов A грамматики разложима и имеет перронов корень равный 1. Целью работы является обобщение результатов, полученных в [2], на случай неограниченного числа классов нетерминалов.

Стохастической КС-грамматикой называется система $G = \langle V_N, V_T, R, s \rangle$, 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 одинаковой левой частью вида

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

где $A_i \in V_N$, $\beta_{ij} \in (V_T \cup V_N)^*$ и p_{ij} — вероятность применения правила r_{ij} , причём $0 < p_{ij} \leq 1$ и $\sum_{j=1}^n p_{ij} = 1$.

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 некоторые обозначения. Будем говорить, что нетерминал A_j непосредственно следует за нетерминалом A_i (и обозначать $A_i \rightarrow A_j$), 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 \rightarrow обозначим \rightarrow_* .

Классом нетерминалов назовём максимальное по включению подмножество $K \in V_N$ такое, что $A_i \rightarrow_* A_j$ 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 $\mathfrak{K} = \{K_1, K_2, \dots, K_m\}$ — множество классов нетерминалов грамматики, $m \geq 2$. 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 её матрица первых моментов A имеет вид

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

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 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, $r = \max_i \{r_i\}$.

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

Рассмотрим всевозможные подцепочки классов $K_\mu, K_{\mu+1}, \dots, K_\nu$. Число классов с номерами из J в такой подцепочке обозначим через $s_{\mu\nu}$. Обозначим, кроме того, $q_j = s_{jm}$.

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

Теорема 1 Пусть $r_j = 1$. Тогда при $t \rightarrow \infty$

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

где $U^{(j)}$ — правый собственный вектор для матрицы A_{ii} из (1), соответствующий r_j , и c_j — некоторая константа.

Теорема 2 Пусть $r_j < 1$. Тогда при $t \rightarrow \infty$

$$P^{(j)}(t) \sim U^{(j)} \cdot \frac{1}{t^{1+(\frac{1}{2})^{q_j-1}}},$$

$$\text{где } U^{(j)} = (E - A_{jj})^{-1} \sum_l A_{jl} U^{(l)} \cdot c_l,$$

где суммирование ведется по всем номерам l классов с $r_l = 1$ таких, что $K_j \prec_* K_l$ и $q_l = q_j - 1$.

Здесь c_l и $U^{(l)}$ имеют тот же смысл, что и в теореме 1.

Теорема 3 Пусть грамматика имеет вид «цепочки», и $r_j = 1$ для любого $j \in \{1, 2, \dots, m\}$. Пусть, кроме этого, в грамматике имеется некоторое правило r_{ij} , такое что $A_i \in K_\nu$. Тогда при $t \rightarrow \infty$ математическое ожидание числа применений правила r_{ij} в деревьях вывода высоты t выражается следующим образом:

$$M_{ij}(t) = \mathcal{M}u_1^{(1)} \cdot t^{(\frac{1}{2})^{m-\nu-1}} (1 + o(1)), \quad (2)$$

где \mathcal{M} — некоторая константа, и $u_1^{(1)}$ — первая компонента правого собственного вектора матрицы A_{11} .

Таким образом, в цепочке, полностью состоящей из критических классов, асимптотика определяется удалённостью класса, в котором может применяться искомое правило (K_ν в обозначениях теоремы), от последнего класса цепочки. Если класс K_ν значительно удалён от конца цепочки, искомое правило применяется реже, так как высота t дерева вывода достигается в большей степени за счёт применения правил из последующих классов. При приближении K_ν к концу цепочки, искомое правило применяется чаще, в силу большей разветвлённости дерева вывода ближе к его листьям, и большего содержания в нём нетерминалов из K_ν .

Теорема 4 Пусть в грамматике, имеющей вид «цепочки» имеется некоторое правило r_{ij} такое, что $A_i \in K_\nu$ и $\nu \in J$. Тогда при $t \rightarrow \infty$ математическое ожидание числа применений правила r_{ij} в деревьях вывода высоты t выражается следующим образом:

$$M_{ij}(t) = \mathcal{M}U_1^{(1)} \cdot t^{(\frac{1}{2})^{m-\nu-1}} (1 + o(1)), \quad (3)$$

где \mathcal{M} — некоторая константа, и

$$U^{(1)} = (E - A_{11})^{-1} A_{12} \cdot \dots \cdot (E - A_{\mu^*-1, \mu^*-1})^{-1} A_{\mu^*-1, \mu^*} u^{(\mu^*)},$$

если K_{μ^*} — класс с наименьшим номером среди классов подцепочки K_1, \dots, K_ν , имеющих перронов корень, равный 1.

Список литературы

- [1] Гантмахер Ф.Р. **Теория матриц**. — 5-е изд., — М.: ФИЗМАТЛИТ, 2010 — 560 с. — ISBN 978-5-9221-0524-8
- [2] Борисов А.Е. Закономерности в словах стохастических контекстно-свободных языков, порождённых грамматиками с двумя классами нетерминальных символов. Вопросы экономного кодирования.