О числе нетерминалов в деревьях вывода разложимой стохастической КС-грамматики

Мартынов Игорь Михайлович

Нижегородский государственный университет им. Н. И. Лобачевского, e-mail: murbidodrus@gmail.com

В работе исследуются вероятностные свойства деревьев вывода высоты t, порождаемых разложимой стохастической КСграмматикой, при $t \to \infty$. Предполагается, что грамматика согласованна, то есть, перронов корень матрицы A первых моментов грамматики не превосходит 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 соответственно, $s \in V_N$ — аксиома, $R = \cup_{i=1}^{k} R_i$, где k — мощность алфавита V_N и R_i — множество правил вывода вида

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

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 $\sum_{j=1}^{n_i} p_{ij} = 1$.

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 $\omega(\alpha) = (r_1, \ldots, r_s)$ правил грамматики (вывод), с помощью которой α выводится из аксиомы s. Выводу слова соответствует дерево вывода [1] d, вероятность $p(d)$ которого определяется как произведение вероятностей правил, образующих вывод: $p(d) = \prod_{k=1}^{s} p(r_k)$.

Грамматика называется *согласованной*, если сумма вероятностей всех конечных деревьев вывода равна 1. Согласованная стохастическая грамматика G задаёт распределение вероятностей на множестве слов порождаемого ею языка $L(G)$. В дальнейшем всюду будем предполагать, что грамматика солгасованна.

По стохастической КС-грамматике строится матрица A первых моментов. Её 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 [2] матрицы A обозначим через r. Известно, что согласованная грамматика имеет перронов корень $r \leq 1$.

Введём некоторые отношения на множестве нетерминальных символов. 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 $A_i \xrightarrow{p_{il}} \alpha_1 A_j \alpha_2$, 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 $K \subseteq V_N$ 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 $A_1 \to A_2$. Рефлексивное транзитивное замыкание отношения $\prec$ обозначим через $\prec_*$. Классы грамматики, за которыми непосредственно не следует ни один класс, будем называть завершающими. Грамматика называется *разложимой*, если она содержит более одного класса, и *неразложимой* в противном случае.

Случай $r < 1$ рассматривался Л. П. Жильцовой (в [3] и других работах). А. Е. Борисов обобщил [4] полученные результаты на случай $r \leq 1$ для грамматики из двух классов.

Пусть $\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$ — множество классов нетерминалов грамматики, $m \geq 2$. Будем полагать, что классы нетерминалов перенумерованы таким образом, что $i \leq j$ для любых $K_i \prec_* K_j$. Заметим, что при этом класс K_1 содержит аксиому s грамматики. Матрица первых моментов A грамматики имеет следующий вид.

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1,n-1} & A_{1,n} \\ 0 & A_{22} & A_{23} & \cdots & A_{2,n-1} & A_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_{n-1,n-1} & A_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 & A_{n,n} \end{pmatrix}$$

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 [2]. Очевидно, $r = \max_i\{r_i\}$. Классы K_i, перроновы корни r_i которых равны 1, будем называть критическими. Остальные классы грамматики будем называть докритическими.

Для каждого класса K_i рассмотрим всевозможные цепочки классов $K_i \prec K_{j_1} \prec K_{j_2} \prec \ldots \prec K_{j_s}$, где класс K_{j_s} — завершающий. Максимум числа критических классов среди $K_i, K_{j_1}, \ldots, K_{j_s}$ по всем таким цепочкам обозначим q_i, а сами такие цепочки будем называть *насыщенными*.

Через $P_i(t)$ обозначим вероятность множества деревьев вывода высоты t, корень которых помечен нетерминалом A_i. Верна следующая теорема.

Теорема 1. *Пусть матрица первых моментов A разложимой КС-грамматики G имеет перронов корень, равный* 1. *Тогда вероятность $P_i(t)$ деревьев высоты t с корнем в A_i имеет вид:*

$$P_i(t) \sim \tilde{c}_i \cdot t^{-1-\left(\frac{1}{2}\right)^{q_l-1}},$$

где $c_i, \tilde{c}_i$ — некоторые константы, $A_i \in K_l$, и $q_l \geq 1$ — максимальное число критических классов в цепочке от K_l до завершающего класса.

Для каждого класса K_i рассмотрим также всевозможные цепочки классов $K_1 \prec K_{j_1} \prec K_{j_2} \prec \ldots \prec K_i$ из начального класса K_1 грамматики в класс K_i. Максимальное число критических классов в такой цепочке обозначим q_i^-. Верна следующая теорема.

Теорема 2. *Пусть матрица первых моментов A разложимой КС-грамматики G имеет перронов корень, не превосходящий* 1. *Тогда математическое число применений правила r_{ij} в случайном дереве вывода высоты t имеет следующий вид:*

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

где p_{ij} — вероятность правила r_{ij}, d_i — некоторая константа, $A_i \in K_l$, и

$$\tilde{q}_l = q_1 - q_l^-.$$

Таким образом, наибольшую асимптотику имеют $M_{ij}(t)$, для которых A_i расположен в последнем критическом классе K_l какой-либо насыщенной цепочки из K_1 в завершающий класс, либо в докритических классах, следующих за K_l. Величина $q_1 - q_l^-$ для таких правил обращается в 0, и $M_{ij}(t)$ имеет асимптотику t^2.

Список литературы

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

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

[3] Жильцова Л. П. Закономерности применения правил грамматики в выводах слов стохастического контекстно-свободного языка // Математические вопросы кибернетики. Вып. 9. М.: Наука, 2000. С. 101-126

[4] Борисов А. Е. Закономерности в словах стохастических контекстно-свободных языков, порождённых грамматиками с двумя классами нетерминальных символов. Вопросы экономного кодирования.