

# О НИЖНЕЙ ОЦЕНКЕ СТОИМОСТИ КОДИРОВАНИЯ ДЛЯ СТОХАСТИЧЕСКОЙ КС-ГРАММАТИКИ, ИМЕЮЩЕЙ ВИД «ЦЕПОЧКИ», В КРИТИЧЕСКОМ СЛУЧАЕ

Игорь Мартынов (Нижний Новгород)

## Введение

При передаче и хранении информации часто возникает необходимость кодирования данных таким образом, чтобы обеспечить наибольшую степень сжатия. Сжатие данных может быть достигнуто использованием статистических данных, таких как частоты появления букв в сообщениях. Если, кроме этого, учитывать структурные свойства языка сообщений, можно дополнительно увеличить эффективность сжатия.

Исследуется источник сообщений, описываемый разложимой стохастической КС-грамматикой с матрицей первых моментов  $A$  специального вида. Рассматривается критический случай, когда перронов корень матрицы  $A$  равен 1. Для таких грамматик рассматривается множество деревьев вывода фиксированной высоты, оценивается энтропия множества слов, порождённых такими деревьями, а также стоимость кодирования.

## 1. Основные определения

*Стохастической КС-грамматикой* [1] называется система  $G = \langle V_T, V_N, R, s \rangle$ , где  $V_T$  и  $V_N$  — конечные множества терминальных и нетерминальных символов (терминалов и нетерминалов) соответственно,  $s \in V_N$  — аксиома,  $R = \bigcup_{i=1}^k R_i$ , где  $k$  — мощность алфавита  $V_N$ , и  $R_i$  — множество правил вида

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

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

$$0 < p_{ij} \leq 1, \quad \sum_{j=1}^{n_i} p_{ij} = 1.$$

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

Каждому слову  $\alpha$  из КС-языка соответствует последовательность правил грамматики (вывод), с помощью которой  $\alpha$  выводится из аксиомы  $s$ . Выводу слова соответствует дерево вывода [2], 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  $r$ .

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  $A_i \xrightarrow{P_{ij}} \alpha_1 A_j \alpha_2$ , 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  $K \subseteq V_N$  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  $\mathcal{K} = \{K_1, K_2, \dots, K_m\}$  — множество классов нетерминалов грамматики,  $m \geq 2$ . Будем полагать, что классы нетерминалов пронумерованы 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

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

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

$$K_1 \prec K_2 \prec \dots \prec K_i \prec \dots \prec K_m.$$

Для каждого класса  $K_i$  матрица  $A_{ii}$  неразложима. Без ограничения общности будем считать, что она строго положительна и неперiodична. Обозначим через  $r_i$  перронов корень [3] матрицы  $A_{ii}$ . Для неразложимой матрицы перронов корень является вещественным и простым. Очевидно,  $r = \max_i \{r_i\}$ .

## 2. Полученные результаты

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

Рассмотрим подцепочку классов

$$K_j \prec K_{j+1} \prec \dots \prec K_m,$$

такую что  $j \in J$ , и  $i \notin J$  для любого  $i > j$ . Обозначим через  $I = \{j, j+1, \dots, m\}$  множество индексов классов этой подцепочки. Тогда верна следующая теорема:

**Теорема 1.** *Энтропия языка  $L^t$ , состоящего из слов, порождаемых в разложимой стохастической КС-грамматике вида «цепочки» с однозначным выводом деревьями высоты  $t$ , выражается формулой*

$$H(L^t) \sim \sum_{i \in I} \sum_{j=1}^{n_i} d_i H(R_i) \cdot t^2,$$

где  $d_i > 0$ ,  $H(R_i) = \sum_{j=1}^{n_i} p_{ij} \log p_{ij}$  — энтропия множества  $R_i$  правил вывода с нетерминалов  $A_i$  в левой части.

Энтропия  $H(L^t)$  имеет асимптотику  $t^2$ , причём константа при  $t^2$  определяется нетерминалами из наиболее удалённого от начала цепочки критического класса, а также классами следующими за ним.

Будем обозначать через  $L^t$  множество слов, порождённых деревьями высоты  $t$ . Через  $f^*$  обозначим кодирование языка  $L^t$ , минимизирующее величину

$$M_t(f) = \sum_{\alpha \in L^t} p_t(\alpha) \cdot |f(\alpha)|,$$

где величины  $p_t(\alpha)$  задают распределение вероятностей на множестве слов  $L^t$ . По определению  $f^*$ , для любого кодирования  $f$  множества слов  $L^t$  справедливо  $M_t(f) \geq M_t(f^*)$ . Тогда верна следующая теорема:

**Теорема 2.** *Пусть матрица  $A$  первых моментов грамматики имеет вид (1) и её перронов корень  $r = 1$ . Тогда стоимость любого кодирования  $f$  языка  $L$ , порождаемого этой грамматикой, удовлетворяет неравенству*

$$C(L, f) \geq C^*(L),$$

где

$$C^*(L) = \frac{\sum_{i \in I_t^+} d_i H(R_i)}{\sum_{i \in I_t^+} d_i L(R_i)},$$

где, в свою очередь,  $H(R_i) = -\sum_{j=1}^{n_i} p_{ij} \log p_{ij}$ ,  $L(R_i) = \sum_{j=1}^{n_i} l_{ij} p_{ij}$ ,  $l_{ij}$  — число терминальных символов в правой части правила  $r_{ij}$ , и  $d_i > 0$  — некоторая константа.

## Список литературы

1. Фу К. Структурные методы в распознавании образов. М.: Мир, 1977
2. Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. Том 1. М.: Мир, 1978
3. Гантмахер Ф. Р. Теория матриц. — 5-е изд., — М.: ФИЗМАТЛИТ, 2010