АЛГОРИТМ ХАФФМЕНА

Н. А. Соловьёва

Кафедра высшей математики

Санкт-Петербургский Государственный Инженерно-экономический Университет

Математический кружок

Код переменной длины

К. Шеннон и Р. Фано предложили конструкцию кода переменной длины, в котором у каждого символа своя длина кодовой последовательности. Вопрос: как определять завершение кода отдельного символа. В качестве решения используется префиксный код.

Префиксный код

Свойство префикса: никакая кодовая последовательность не является началом другой кодовой последовательности.

Код, обладающий таким свойством, называется **префиксным кодом.**

Характеристика кода

Предположим, что кодируемые символы появляются в тексте независимо. Обозначим через p_i вероятность появления i-го символа, через s_i — длину его кодовой последовательности. Тогда математическое ожидание длины кодовой комбинации случайно выбранного символа равно

$$\sigma = \sum_{i=1}^{m} p_i s_i \,.$$

Экстремальная задача

Неравенство Крафта:

$$\sum_{i=1}^{m} 2^{-s_i} \leq 1.$$

Задача о префиксном коде:

минимизировать математическое ожидание σ по всем наборам длин $\{s_i\}$, удовлетворяющим неравеству Крафта.

Код, для которого σ минимально, называется **оптимальным.**

Свойства оптимального кода

Лемма 1. Пусть $\{p_i\}$ — набор вероятностей символов и $\{s_i\}$— длины оптимальных кодовых комбинаций. Если $p_1 \geq p_2 \geq \cdots \geq p_n$, то $s_1 \leq s_2 \leq \cdots \leq s_n$.

Лемма 2. В обозначениях и предположении леммы 1 две самые длинные кодовые комбинации имеют одинаковую длину, то есть $s_{n-1} = s_n$.

Свойства оптимального кода

Лемма 3. Рассмотрим наравне с исходной задачей P сокращённую задачу P', которая получается объединением двух самых редких символов в один символ. Минимальное значение целевой функции в задаче P' отличается от значения в задаче P на $p_{n-1} + p_n$, а оптимальный кодовый набор для задачи P получается из решения задачи P' удлинением на один бит кода объединённого символа.

Алгоритм Хаффмена

Алгоритм Хаффмена:

если в алфавите два символа, то нужно закодировать их 0 и 1, а если больше, то соединить два самых редких символа в один новый символ, решить получившуюся задачу и вновь разделить этот новый символ, приписав 0 и 1 к его кодовой последовательности.

Пример: алгоритм Хаффмена

Пусть алфавит состоит из пяти символов a, b, c, d, e с вероятностями появления 0.37, 0.22, 0.16, 0.14, 0.11 соответственно.

$$0.37(a) \quad 0.22(b) \quad 0.16(c) \quad \underbrace{0.14(d) \quad 0.11(e)}$$

$$0.37(a) \quad 0.25(de) \quad \underbrace{0.22(b) \quad 0.16(c)}$$

$$0.38(bc) \quad \underbrace{0.37(a) \quad 0.25(de)}$$

$$0.62(ade) \quad 0.38(bc)$$

Пример: обратный ход

Выполним обратный ход алгоритма:

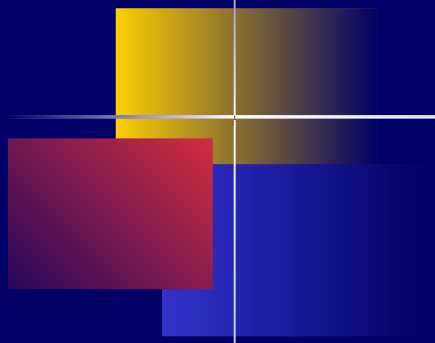
$ade\ 0$	$a\ 00$	$a\ 00$
	$de\ 01$	$d\ 010$
		$e\ 011$
$bc\ 1$	$b\ 10$	$b\ 10$
	$c\ 11$	$c\ 11$

Задачи

1. Представить доказательства лемм 1-3.
2. Написать программу, реализующую алгоритм Хаффмена.
3. Оценить на примерах коэффициент сжатия по алгоритму Хаффмена.

Список литературы

И. В. Романовский. *Дискретный анализ.* СПб: Невский Диалект, БХВ–Петербург, 2004.



Спасибо за внимание!