Одно обобщение алгоритма Шеннона-Фано для кодирования дискретных множеств сообщений

Герасимов М. А., Санкт-Петербургский государственный университет,
ge@star.math.spbu.ru

Аннотация

Предлагается обобщение алгоритма Шеннона-Фано, которое позволяет не только находить оптимальные коды для заданного конечного множества сообщений за полиномиальное время, но и дает возможность кодировать сообщения приближенно, в зависимости от имеющихся ресурсов (времени и памяти). Оценки сложности сделаны для детерминированной одноленточной машины Тьюринга с входной и выходной лентой.

Машина Тьюринга

Для оценки сложности алгоритмов рассматривается одноленточная одноголовочная машина Тьюринга с входной и выходной лентой для записи результата. Предполагается, что входные данные записываются на входной ленте, обрабатываются на рабочей ленте и результат записывается на выходной ленте. При работе машины Тьюринга используется алфавит, состоящий из четырех символов {#,b,0,1}. Результатом работы алгоритма считается битовая последовательность, кодирующая исходное множество сообщений, и соответствующее дерево кодирования, позволяющее однозначно восстановить исходную последовательность. В дальнейшем будем считать, что входные данные (натуральные числа) записаны в виде битовой последовательности на входной ленте между маркерами '#'. В качестве разделителя входных битовых последовательностей используется пустой символ 'b'. Считывание второго маркера означает конец цепочки входных данных. Входная лента позволяет считывать входные данные произвольное количество раз. Выходная лента позволяет только записать результат вычисления в виде последовательности символов рабочего алфавита. Каждый символ выходной цепочки записывается только один раз и больше не изменяется.

Алгоритм Шеннона - Фано

Рассматривается общий вид алгоритма Шеннона-Фано, а именно, предполагается наличие некоторого множества сообщений X = $\{x_1, , x_M\}$ и

некоторого распределения p, как функции из множества X во множество рациональных чисел от 0 до 1, обладающей тем свойством, что $\sum_{i=1}^{M} p(x_i) = 1$. Для простоты будем предполагать наличие алфавита A, состоящего из 2-х символов {0,1}. Результатом работы алгоритма будет считаться набор цепочек, кодирующих сообщения множества X в алфавите A, таким образом, что существует обратное однозначное отображение, восстанавливающее по цепочке из алфавита A соответствующее сообщение из множества X. Оптимальность понимается в обычном смысле, т.е. средняя длина полученных кодирующих цепочек (математическое ожидание) отличается от энтропии множества X, при заданном распределении p, на минимальную величину, обычно близкую или равную 0.

Предполагается также, что алгоритм производит следующие шаги:

Шаг 1.Строит дерево кодирования для множества X по следующим правилам:
Шаг 1.1.Если |X|=1, то алгоритм останавливается, помещая это множество в лист дерева кодирования.
Шаг 1.2.Если |X|>1, то исходное множество разбивается на два непустых подмножества X_1, X_2 таким образом, что сумма вероятностей сообщений одного множества отличается от суммы вероятностей сообщений другого множества не более чем на некоторое $\delta > 0$, возможно равное 0. К множествам X_1, X_2 рекурсивно применяется шаг 1.1.
Шаг 2.Используя построенное дерево, листьями которого являются одноэлементные множества, строится код $: X- > A^*$ по естественным образом: левой ветви дерева соответствует символ '0', правой ветви дерева соответствует '1'. Путь из корня дерева в лист, соответствующий элементу x будет определять код этого элемента.
Шаг 3.Полученные коды сообщений множества X записываются на выходную ленту машины Тьюринга.
Данный алгоритм завершает свою работу не более чем за |X| шагов, при любом множестве X, поскольку на шаге 1.2 мощность любого из получаемых подмножеств X_1, X_2 меньше мощности X. Полученный в результате работы алгоритма код может не быть оптимальным и может иметь относительную погрешность, зависящую от выбора δ.
Теорема 1. *Для любого множества X существует $\delta(X) > 0$, которое применимо к любому подмножеству множества X на шаге 1.2. обобщенного алгоритма Шеннона-Фано.*
Теорема 2. *Существует полиномиальный по времени алгоритм, кодирующий любое множество сообщений X цепочками двоичного алфавита $A =$*

$\{0,1\}$ *с любым распределением p, обобщенным алгоритмом Шеннона-Фано с* $\delta(X) = max_{x \in X} p(x) - min_{x \in X} p(x)$.

Замечание 1. *Данное определение, теорему 1, и теорему 2 можно обобщить до случая, когда мощность алфавита |A|>2.*

Замечание 2. *Обобщенный алгоритм Шеннона-Фано работает и в тех случаях, когда существуют* x_1, x_2 *из X, для которых* $p(x_1) = p(x_2)$.

Литература

[1] Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982.

[2] Fischetti M., Martello S. Wost–case analysis of the differencing method for the partition problem.// Math. Programming. 1987. V 37, N 1 pp 117-120.

[3] Минский М. Вычисления и автоматы. — М., 1971.

[4] Huffman. D. A method for construction of minimum redundancy codes. Proceeding of IRE, 40(9):1098–1101, 1952.

[5] Horowits E., Sahni S. Fundamentals of Computer Algorithms. Computer Science Press, 1978.

[6] Shannon C. E. A Mathematical Theory of Communication, Bell System Technical Journal, vol. 27, July 1948, pp. 373–423.