УДК 621.391

А. В. Аксенов – магистрант кафедры вычислительных систем и сетей
А. Ю. Козловский (ст. преп.) – научный руководитель

РАЗРАБОТКА АДАПТИВНОГО АЛГОРИТМА ЭНТРОПИЙНОГО КОДИРОВАНИЯ

Экономное кодирование (сжатие без потерь) подразумевает создание эффективного представления некоторой информационной выборки. Под ней будем подразумевать выборку источника дискретной информации с конечным алфавитом. Сжатие становится возможным благодаря наличию у информации определенных свойств, которые можно использовать, создав как можно более точное описание источника.

Существует два основных подхода к описанию информационных источников.

1. Комбинаторный подход, где символы источника рассматриваются группами – сообщениями, причем источник может посылать только равновероятные сообщения из определенного множества, которое является описанием. К положительным качествам такого подхода можно отнести высокое быстродействие алгоритмов, использующих такой способ, из-за простоты описания источника. При этом описание является неточным, поскольку реальные источники могут посылать очень большое количество сообщений, причем вероятности появления этих сообщений чаще всего различны.

2. Вероятностный подход, где вычисляются вероятностные оценки появления символов на выходе источника информации. Данный метод более вычислительно сложен, но описание свойств источника, порождаемое им, является более точным.

Клод Шеннон ввел понятие энтропии как меры описания неопределенности информации. Если нам дан информационный источник с вероятностным распределением появления символов $\{p_i\}_{i=1}^{N}$, то его энтропия вычисляется как

$$H(p_1, p_2, ..., p_N) = -\sum_{i=1}^{N} p_i \log_m p_i ,$$

где m – основание системы представления информации [1].

При этом считается, что источник находится в некотором состоянии. Переход между состояниями, сопровождающийся изменением вероятностного распределения порождения символов, происходит при выделении очередного символа. Состояние источника полностью определяется порожденной им информационной выборкой.

В целом для информационного источника S с T состояниями, для которого полностью определена матрица переходных вероятностей

$$P = \{p_{ij}\},$$

где p_{ij} - вероятность перехода из состояния i в состояние j, и если существует

$$P_\infty = \lim_{k\to\infty} P^k ,$$

величина энтропии вычисляется по формуле

$$H(S) = -\sum_{i=1}^{T} p_i \sum_{j=1}^{T} p_{ij} \log_m p_{ij} .$$

Множество состояний источника в совокупности с переходными вероятностями – марковская модель источника. Для определения энтропии необходимо построить марковскую модель максимально близко к модели источника, что на практике затруднительно, т.к. априорно параметры источников неизвестны. На практике вычисление энтропии производится последовательным анализом переходов между состояниями источника. При этом величина энтропии усредняется по времени.

Рассмотрим вероятностный источник, который последовательно находится в n состояниях $s_1, s_2, ..., s_n$. Состоянию s_i соответствует распределение вероятностей появления на выходе символов $\{p_j^{(s_i)}\}_{j=1}^{N}$. Энтропия источника при этом может быть получена по формуле

$$H'(S) = \frac{-\sum_{i=1}^{N}\sum_{j=1}^{n} p_j^{(s_i)} \log_m p_j^{(s_i)}}{n}.$$

Тем не менее, в реальных ситуациях мы не обладаем априорной информацией об источнике вообще, поэтому априорные распределения вероятностей $\{p_j^{(s_i)}\}_{j=1}^{N}$ должны быть исключены из расчетов и заменены на их эмпирические оценки. Формула вычисления чисто эмпирической энтропии выглядит следующим образом:

$$H''(S) = \frac{-\sum_{i=1}^{N}\sum_{j=1}^{n} r_j^{(s_i)} \log_m r_j^{(s_i)}}{n}, \quad (1)$$

Где $r_j^{(s_i)}$ - эмпирическая оценка вероятности порождения i-го символа источником, находящимся в j-ом состоянии (породившим j-ю информационную выборку).

Энтропия важна, т.к. является границей эффективности кодирования выборки информационного источника. При этом многие реальные источники по природе не являются вероятностными, т.е. не имеющие некоторых четких состояний, поэтому особо важную роль играет именно чисто эмпирическая энтропия, когда за состояние источника принимается факт порождения им некоторой информационной выборки.

Как видно из формулы (1), построение марковской модели информационного источника может быть сведено к построению вероятностной модели для последовательного вычисления эмпирических оценок порождения некоего символа источником, породившим данную информационную выборку.

Можно рассчитать оптимальный вклад символа в длину кода [2], он составляет

$$x_a = -\log_m p_a$$

Для генерации кодов при использовании вероятностных методов кодирования могут использоваться префиксные системы кодов (код Шеннона-Фано, код Хаффмана, код Голомба и другие). Однако все они обладают рядом недостатков:

- большая требовательность к вычислительным ресурсам (экспоненциальный рост количества префиксных кодов при увеличении длины сообщений);
- генерация кода только после обработки всей информации;
- редкая достижимость оптимальной длины кода символа (необходимость целой длины префиксного кода).

Последний недостаток проиллюстрирован на рисунке 1. Метод Хаффмана, который является оптимальным среди префиксных кодов, достигает оптимальной длины кода отдельного символа лишь для символов, вероятность порождения которых составляет

$$p_a = m^{-n}, n = 1,2,3,...$$

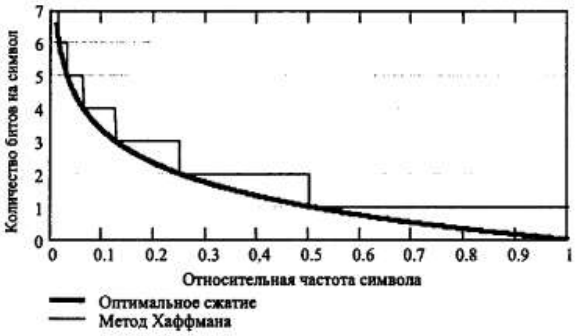

Рис. 1. Сравнение метода Хаффмана с оптимальным сжатием для m=2

Всех этих недостатков лишено арифметическое кодирование. Его следует признать оптимальным методом энтропийного кодирования.

Возможны несколько видов кодеров, использующих арифметическое кодирование. Наиболее простым случаем являются неадаптивные кодеры, которые моделируют стационарный источник (источник, не имеющий переходов между состояниями, у которого вероятностное распределение символов, а, следовательно, и энтропия постоянны по времени). Помимо неточного и негибкого описания источника такие методы имеют чисто технические недостатки: стационарная модель должна либо задаваться априорно, что недопустимо, либо синтезироваться на основе анализа источника до начала кодирования, и при этом включаться в код, поскольку декодер не будет располагать исходной информацией для синтеза модели.

Более удобен адаптивный метод, характеризующийся следующими принципами:

1) Кодер и декодер начинают с моделями в некотором оговоренном «нулевом» состоянии.

2) Кодер и декодер изменяют вероятностную модель эквивалентно.

3) В процессе кодирования используется только информация, доступная на этапе декодирования (эмпирическая оценка вероятностного распределения порождаемых символов).

4) Производится масштабирование статистики модели (учет локальности).

Большинство практических реализаций адаптивных энтропийных кодеров строятся по схеме, отраженной на рисунке 2. При этом особенно важно, что текущий символ кодируется на основании состояния модели, ему предшествующего, чтобы сделать возможным корректное декодирование. Это влечет необходимость в последующем обновлении модели.

Рис. 2. Обобщенная схема реализации адаптивного энтропийного кодера

Кроме того, на практике модель оперирует не эмпирическими вероятностями символов p_i, а их относительными частотами $r_i = \frac{n_i}{n}$, где n_i - число порождений i-го символа в данном состоянии, n - число порождений всех символов в данном состоянии. Отсюда, в частности, следует проблема «нулевой вероятности», когда i-й символ ни разу не был порожден этим состоянием и, следовательно, не может быть закодирован при первом порождении. Практические реализации кодеров используют различные методы борьбы с этим явлением. [3]

По методу идентификации состояний источника адаптивные кодеры делятся на 2 класса:

- контекстно-свободные, когда каждый новый порожденный символ переводит источник в новое, ранее не встречавшееся состояние, для которого вычисляются эмпирические оценки на основании предыдущего состояния с учетом предыдущего порожденного символа;

- контекстно-зависимые, когда состояние источника связывается с некоторой последовательностью символов, обладающей вероятностными характеристиками – контекстом. Как правило, в качестве контекста рассматривается последовательность символов некоторой длины, порожденная источником перед переходом в текущее состояние. При этом достигается более высокая точность эмпирической оценки вероятностного распределения порождения символов в текущем состоянии.

Библиографический список

1. *Shannon C.E.* A mathematical Theory Of Communication, 1948
2. *Семенюк В.В.* Вероятностные методы кодирования видеоинформации, 2004
3. *Bunton S.* Online Stohastic Processes in Data Compression, 1996