UDC 004.62

Kutsenko A.A. Comparative analysis of lossless hyperspectral compression methods

Сравнительный анализ методов сжатия гиперспектральных данных без потерь

Kutsenko Alexander Alekseevich
Bauman Moscow State Technical University, Russia, Moscow
Куценко Александр Алексеевич
Федеральное государственное автономное образовательное учреждение высшего образования «Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)», РФ, г. Москва

Abstract. *The article analyzes lossless hyperspectral data compression methods, including tANS and Huffman coding. It is shown that tANS combines the efficiency of arithmetic coding with the speed of Huffman coding, achieving compression rates close to the Shannon limit. Due to its implementation simplicity and small table size (1–2 KB), the method is suitable for embedded image processing systems.*

Keywords: *lossless compression; entropy coding; arithmetic coding; Huffman coding; tANS.*

Аннотация. *В статье проведён анализ методов сжатия данных без потерь, включая tANS и кодирование Хаффмана. Показано, что tANS сочетает эффективность арифметического сжатия и быстродействие Хаффмана, обеспечивая степень сжатия, близкую к пределу Шеннона. Благодаря простоте реализации и малому расходу памяти, метод применим во встраиваемых системах для сжатия данных.*

Ключевые слова: *сжатие данных без потерь; энтропийное кодирование; арифметическое кодирование; кодирование Хаффмана; tANS.*

Рецензент: Сагитов Рамиль Фаргатович - кандидат технических наук, доцент. Заместитель директора, главный научный сотрудник. ООО «Научно-исследовательский проектный институт «Промышленное и гражданское строительство»

Сжатие данных – это процесс представления исходной информации в более компактной форме при возможности её точного восстановления либо при допустимом уровне потерь информации. С точки зрения теории информации, предел эффективности сжатия зависит от статистических свойств источника данных и определяется его энтропией, вычисляемой по формуле Шеннона [1]:

$$H = -\sum_{i} \quad p(s_i)p(s_i)\,,$$

где $p(s_i)$ – вероятность появления символа s_i из множества всех символов алфавита $A = \{s_i\}$. Следовательно, для любого источника данных с энтропией H невозможно осуществить сжатие без потерь сильнее, чем до H бит на символ. Оптимальный код использует каждый бит информации максимально эффективно и характеризуется средней длиной, равной энтропии источника.

В целом эффективность сжатия может быть достигнута двумя подходами:

1) использованием алгоритмов, которые кодируют элементы с числом бит, пропорциональным вероятности их появления (энтропийное кодирование);
2) применением моделей данных, позволяющих отбрасывать малозначимые компоненты, несущественные для восприятия или анализа (моделирующее сжатие).

Алгоритмы сжатия разделяют на два класса: с потерями и без потерь. Методы сжатия с потерями (второй тип в списке выше) предполагают использование модели источника, аппроксимирующей его спектральные или статистичесие характеристики. Компоненты, не описываемые моделью, затем отбрасываются как малозначимые и кодируются методами сжатия без потерь, что позволяет достичь высокой эффективности сжатия за счёт частичной потери информации. Методы без потерь, напротив, обеспечивают восстановление исходных данных бит-в-бит, но обычно достигают меньших коэффициентов сжатия.

В компрессорах данных используются методы двух типов: методы предобработки данных и методы сжатия. Методы преобразования данных направлены на устранение избыточности и эффективное кодирование информации. Они используются для изменения статистических параметров сжимаемых данных с целью оптимального кодирования энтропийным кодировщиком. Методы сжатия выполняют кодирование передаваемых символов кодами разной длины, то есть осуществляют непосредственно сжатие данных. Для некоторых методов известен закон распределения символов, при котором данный метод показывает лучшую эффективность. Другие же используют оценку вероятностей появления символов и могут оптимально кодировать данные с разным распределением.

К простейшему методу энтропийного кодирования относятся коды Голомба-Райса. Кодирование Голомба-Райса представляет собой эффективный метод сжатия данных от целочисленных источников с геометрическим распределением, то есть в данных большая плотность значений около нуля и она уменьшается с увеличением этих значений. Данный метод подходит в случае, когда малые значения встречаются существенно чаще больших [1]. Идея метода заключается в разделении исходного числа n на две части:

1) частное $q = n\ div\ m,\ m = 2^k$,
2) остаток $r = n\ mod\ m$.

Частное q представляется унарным кодом, остаток r прямым кодом переменной длины. Преимуществом данного подхода является простота реализации (особенно благодаря быстрым битовым операциям), высокая эффективность для источников с

геометрическим распределением данных, адаптивность (параметр k можно менять для блоков данных по отдельности), а также отсутствие дополнительных расходов по памяти. К недостаткам можно отнести рост длины кода при больших n, неэффективность для равномерных распределений, а также чувствительность к выбору k (влияет на качество сжатия).

Следующим по простоте реализации является кодирование Хаффмана. Кодирование с использованием кода Хаффмана представляет собой метод энтропийного сжатия данных, основанный на частотном распределении символов входного алфавита. Алгоритм строит префиксное двоичное дерево, в котором часто встречающимся символам сопоставляются более короткие кодовые последовательности, а редко встречающимся – длинные.

Для построения оптимального набора кодов требуется получить статистику частоты символов, что может быть выполнено заранее при первом проходе по данным (статический вариант) или динамически – в процессе кодирования и декодирования (в этом случае используется адаптивный алгоритм Хаффмана).

Преимуществом метода является то, что генерируемые коды Хаффмана являются префиксными, то есть никакой символ не является началом другого. За счёт этого упрощается декодирование: достаточно последовательно читать символы из закодированного потока до тех пор, пока накопленная часть не совпадет с одним из имеющихся кодов. Важным является и то, что кодирование Хаффмана в худшем случае не увеличивает длину сообщения по сравнению с длиной входного сообщения.

К недостаткам можно отнести необходимость двойного прохода по последовательности (для подсчета статистики и непосредственно для кодирования). Этот недостаток можно исправить, воспользовавшись адаптивным вариантом метода, который кодирует символы на лету. Также в этом случае отпадает необходимость в хранении таблицы символов.

Другим вариантом энтропийного сжатия является арифметическое кодирование. Арифметическое кодирование – это метод сжатия без потерь, при котором все сообщение (последовательность символов) представляется в виде одного вещественного числа в интервале [0, 1). В отличие от кода Хаффмана, который сопоставляет каждому символу отдельный код фиксированной длины, арифметическое кодирование описывает вероятность появления всей последовательности, постепенно сужая числовой интервал по мере обработки символов. Такой подход позволяет достигать эффективности, близкой к теоретическому энтропийному пределу, даже для небольших алфавитов и статистически зависимых последовательностей [1].

Основная идея метода заключается в итерационном уточнении интервала [0,1) на

основании вероятностной модели символов. Для каждого символа исходного алфавита априорно известна вероятность его появления $p(s_i)$. Для каждого символа вычисляется верхняя и нижняя границы его подынтервала внутри текущего рабочего интервала, смещенной от его начала на накопленную сумму вероятностей предыдущих символов. Далее, при последовательной обработке символов рабочий интервал сужается до подынтервала, соответствующего очередному символу. В итоге, любое число из финального интервала (чаще всего выбирают среднее значение из интервала) однозначно представляет все исходное сообщение.

Преимущество метода заключается в том, что распределение вероятностей описывается более точно. Происходит это из-за того, что коды Хаффмана ограничены степенью двойки, тогда как арифметическое кодирование использует весь диапазон вещественных чисел для этого.

Недостатком метода является его вычислительная сложность. Обработка каждого символа требует поиска нужного диапазона в таблице символов (задача облегчается, если значения отсортированы по убыванию), а также сложными операциями вещественного умножения (по 2 операции на обработку одного символа). Помимо этого итоговое состояние кодера, то есть финальное число, достаточно быстро достигает предела точности вещественной арифметики. По этой причине используется операция увеличения текущего рабочего диапазона до предыдущего (большего) тогда, когда часть бит текущего состояния фиксируется. Эта операция требует дополнительных вещественных вычислений.

В [2] приведено сравнение арифметического кодера и кодировщика Хаффмана. В результате, использование арифметического кодера дает преимущество по степени сжатия от 6 % до 100 % по сравнению с кодировщиком Хаффмана. Однако время работы такого кодера дольше в 2-4 раза.

Современным продолжением развития арифметического кодировщика являются методы асимметричных систем счисления (англ. Asymmetric Numeral Systems, ANS). ANS алгоритмы объединяют эффективность арифметического кодирования со скоростью кодирования Хаффмана [3].

Основная идея ANS методов заключается в представлении последовательности символов в виде единственного целого числа – состояния, которое увеличивается по мере обработки последовательности. Идея схожа с арифметическим кодером, но вместо двух вещественных чисел используется одно целое, благодаря чему количество операций умножения при обработке каждого символа уменьшается вдвое (с двух до одной). Известны варианты реализации ANS алгоритмов для разных алфавитов: qABS – для бинарных алфавитов, rANS – потоковый компрессор для больших алфавитов, tANS –

версия rANS, использующая машину состояний для описания процесса кодирования.

Особенного внимания заслуживает табличный tANS поскольку в нем работа метода ограничивается диапазоном I_S с помощью таблицы переходов в виде конечного автомата. По сути, формирование этой таблицы переходов является единственной ресурсоемкой операцией. После того как таблица сформирована, вся работа кодера состоит из переходов из одного состояние в другое с выдачей бит в результирующий поток. Благодаря последнему свойству, tANS метод применим даже во встраиваемых системах для сжатия гиперспектральных изображений.

В работе [4] перечислены основные преимущества и недостатки tANS метода. К преимуществам можно отнести следующее:

1) высокая скорость работы – операции просты и выполняются через таблицы и целочисленную арифметику, подходящую для SIMD и аппаратных реализаций;
2) эффективность – коэффициент сжатия близок к энтропийному пределу (зависит от размера таблицы);
3) компактность – таблицы обычно занимают не более 1-2 Кбайт.

Так, для таблицы размера 256 tANS работает в 2 раза быстрее Хаффмана при схожей с арифметическим кодером степенью сжатия [4]. Стоит отметить, что точность приближения к энтропийной степени сжатия зависит от размера таблицы. Оптимальных значений можно достичь, если взять таблицу в 4 или 8 раз больше базовой (включающей все символы алфавита 1 раз) [5].

К недостаткам можно отнести следующее:

1) необходимость предварительного построения таблиц на основе статистики;
2) сложность реализации по сравнению с простыми кодами (например, Хаффмана);
3) эффективность может снижаться на очень коротких блоках данных.

Таким образом, tANS представляет собой современный гибридный метод энтропийного сжатия, сочетающий эффективность арифметического подхода и быстродействие кода Хаффмана. Он реализует числовое представление данных через переходы между состояниями с фиксированной битовой ёмкостью, обеспечивая одновременно высокую скорость и теоретическую эффективность, близкую к пределу Шеннона. В случае если вычислительные ресурсы сильно ограничены, можно воспользоваться кодами Хаффмана, однако tANS метод показывает существенно более высокую степень сжатия.

References

1. Ватолин Д., Ратушняк А., Смирнов М., Юкин В. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. – М.ДИАЛОГ-МИФИ, 2003. – С. 383.

2. Shahbahrami A. Evaluation of Huffman and Arithmetic Algorithms for Multimedia Compression Standards // International Journal of Computer Science, Engineering and Applications (IJCSEA). — 2011. — С. 11.

3. Brian K. Lossless Compression with Asymmetric Numeral Systems [Электронный ресурс]. — 2020. — URL: https://bjlkeng.io/posts/lossless-compression-with-asymmetric-numeral-systems/ (дата обращения 10.02.2026).

4. Duda J. Asymmetric numeral systems: entropy coding combining speed of Huffman coding with compression rate of arithmetic coding // arXiv preprint arXiv:1311.2540. — 2013.

5. Yamamoto H. Encoding and Decoding Algorithms of ANS Variants and Evaluation of Their Average Code Lengths // arXiv preprint arXiv:2408.07322. – 2024.