Колмогоровская сложность

С. Косухин *

23 ноября 2005 г.

План лекции

1. Краткая история вопроса
2. Немного теории
3. Способы применения

1 Краткая история вопроса

Понятие колмогоровской сложности (алгоритмической энтропии) появилось в 1960-е годы на стыке теории алгоритмов, теории информации и теории вероятностей. Идея Колмогорова состояла в том, чтобы измерять количество информации, заключччной в индивидуальных конечных объектах. Схожие идеи высказывал Р. Соломонов, пытаясь ввести понятие априорной вероятностиь. В статье 1969 года американский математик Г. Чейтин дал то же определение алгоритмической сложности, что и Колмогоров. Основные же свойства колмогоровской сложности были исследованы в 1970-е годы.

2 Немного теории

2.1 Что такое колмогоровская сложность?

Допустим у нас есть файл. Сожмем его каким-нибудь архиватором. Получим сжатую версию файла, по которой сможем восстановить исходный. В первом приближении колмогоровскую сложность файла можно определить как длину его сжатой версии. Тем самым файл, имеющий регулярную структуру и хорошо сжимаемый, имеет малую колмогоровскую сложность. Это описание нуждается в уточнениях:

- Вместо файлов будем рассматривать двоичные слова.

*Законспектировал А. Яковлев.

- Программы сжатия нет — мы рассматриваем лишь программу восстановления.

Назовем декомпрессором произвольный алгоритм (программу), который получает на вход двоичные слова и выдает на выход также двоичные слова. Если декомпрессор D на входе x дает y, мы пишем $D(x) = y$ и говорим, что x является описанием y при данном D. Декомпрессоры мы будем также называть способами описания. Пусть фиксирован некоторый способ описания D. Для данного слова x рассмотрим все его описания, то есть все слова y, для которых $D(y)$ определено и равно x. Длину самого короткого из них и называют колмогоровской сложностью слова x при данном способе описания D:

$$KS_D(x) = \min\{l(y)|D(y) = x\} \tag{1}$$

Здесь и далее $l(y)$ обозначает длину слова y. Способ D_1 не хуже способа D_2, если при некотором c и всех x выполнено:

$$KS_{D_1}(x) \leq KS_{D_2}(x) + c \tag{2}$$

Пусть даны два произвольных способа описания D_1 и D_2. Покажем, что существует способ описания D, который не хуже их обоих. Положим

$$D(0y) = D_1(y), D(1y) = D_2(y) \tag{3}$$

Ясно, что если y является описанием x при D_1 (или D_2), то $0y$ (соответственно $1y$) является описанием x при D, и это описание лишь на бит длиннее. Поэтому

$$KS_D(x) \leq KS_{D_1}(x) + 1, KS_D(x) \leq KS_{D_2}(x) + 1 \tag{4}$$

При всех x, так что способ D не хуже обоих способов D_1 и D_2.

Теорема (Соломонов — Колмогоров). Существует способ описания D, который не хуже любого другого: для всякого способа описания D‘ найдется такая константа c, что

$$KS_D(x) \leq KS_{D'}(x) + c \tag{5}$$

Доказательство. Определим новый способ описания D, положив

$$D(py) = p(y)$$

где y — любое двоичное слово, а p — произвольная программа (также двоичное слово), по которой можно определить, где она заканчивается. D читает слово слева направо, выделяя из него начало-программу, и применяет еч к остатку входа y, после чего выдает результат. Покажем, что D не хуже любого другого способа описания P. Пусть p — программа, соответствующая способу описания P, причем записанная в выбранной форме. Если слово y является описанием слова x относительно P, то py будет описанием x относительно D. Поэтому при переходе от P к D длина описания увеличится

не более чем на длину p, зависящую только от выбора способа описания P, но не от x.

Способ описания, обладающий указанным в теореме свойством, называют **оптимальным**. Замена оптимального способа на другой оптимальный способ приводит к изменению сложности не более чем на константу. Выбирая тот или иной оптимальный способ, можно сделать сложность конкретного слова любой, поэтому не имеет смысла говорить о колмогоровской сложности без задания оптимального описания.

2.2 Сложность и информация

Колмогоровскую сложность слова x можно также назвать количеством информации в слове x. К примеру, в слове 0000000000000, похоже, маловато информации в то время как в 0110101110110 ее уже куда больше! (осмысленность в данном случае не интересует).

Если слово x имеет сложность k, мы говорим, что x содержит k битов информации. Естественно ожидать, что число битов информации в слове не превосходит его длины.

Теорема. Существует такая константа c, что для любого слова x выполнено $KS(x) \leq l(x) + c$

Доказательство. Если $P(y) = y$ при всех y, то $KS_P(x) = l(x)$. В силу оптимальности найдется такое D, что $KS_D(x) \leq KS_P(x) + c = l(x) + c$ при всех x.

Из утверждения теоремы вытекает, что колмогоровская сложность любого слова конечна.

Теорема. Для любого алгоритма A существует такая константа c, что для всех x, где определено $A(x)$, выполнено $KS(A(x)) \leq KS(x) + c$

Доказательство. Пусть D — оптимальный декомпрессор, используемый при определении колмогоровской сложности. Рассмотрим другой декомпрессор D‘:

$$D`(y) = A(D(y)) \tag{6}$$

Если p является описанием некоторого x относительно D, то p является описанием $A(x)$ относительно D‘. Взяв в качестве p кратчайшее описание x относительно D, находим, что

$$KS_{D`}(A(x)) \leq l(p) = KS_D(x) = KS(x) \tag{7}$$

а в силу оптимальности D

$$KS(A(x)) = KS_D(A(x)) \leq KS_D(A(x)) + c \leq KS(x) + c \tag{8}$$

при некотором c и при всех x.

Эта теорема гарантирует, что количество информации не зависит от кодировки.

Пусть x и y — два слова. Соединим их в одно, приписав y к x справа. Сколько битов информации будет иметь полученное слово?

Теорема (о сложности конкатенации). Существует такая константа c, что при любых x и y выполнено неравенство:

$$KS(xy) \leq KS(x) + 2\log KS(x) + KS(y) + c \tag{9}$$

Доказательство. Докажем теорему без ослабляющего добавочного члена $2\log KS(x)$. Пусть D — оптимальный способ описания. Рассмотрим другой способ описания D‘. Именно, если $D(p) = x$ и $D(q) = y$, будем считать pq описанием слова xy, то есть положим D‘$(pq) = xy$. Тогда $KS_D(xy) \leq l(pq) = l(p) + l(q)$. Если в качестве p и q взять кратчайшие описания, то получится

$$KS_D(xy) \leq KS_D(x) + KS_D(y)$$

Остается воспользоваться оптимальностью D и перейти от D‘ к D, где и возникает константа c.

Вопрос. Что неверно в рассуждении?

Ответ. Определение D‘ некорректно: мы полагаем D‘$(pq) = D(p)D(q)$, но D‘ не имеет средств разделить p и q. Вполне может оказаться, что есть два разбиения слова на части, дающие разные результаты: $p_1q_1 = p_2q_2$, но $D(p_1)D(q_1) \leq D(p_2)D(q_2)$. Исправим ошибку: перед словом p напишем его длину. Обозначим через $bin(k)$ двоичную запись числа k, а через x результат удвоения каждого бита в слове x. Теперь положим D‘$(bin(l(p))01pq) = D(p)D(q)$ Это определение корректно. Теперь величина $KS_{D\text{‘}}(xy)$ оценивается числом $2l(bin(|p|)) + 2 + l(p) + l(q)$ Двоичная запись числа $l(p)$ имеет длину не больше $\log_2 l(p) + 1$, поэтому D‘-описание слова xy имеет длину не более $2\log_2 l(p) + 4 + l(p) + l(q)$, откуда и вытекает утверждение теоремы.

Насколько неравенство о сложности конкатенации строк близко к равенству? Пусть x и y содержат много общего. Например, при $x = y$ мы имеем

$$KS(xy) = KS(xx) = KS(x) + O(1),$$

поскольку xx алгоритмически получается из x и обратно.

Вопрос. Пусть задан некоторый текст. У какого перевода нулевая колмогоровская сложность?

Ответ. К примеру, у тождественного либо у машинного.

Определим понятие количества информации, которая содержится в x, но не содержится в y. Эту величину называют также **условной колмогоровской сложностью** x при условии y и обозначают $KS(x|y)$.

Разность $KS(x) - KS(x|y)$ естественно назвать **количеством информации** об x в y. Тем самым приобретает смысл вопрос о том, сколько новой информации в ДНК собаки по сравнению с ДНК волка: если w — двоичное слово, кодирующее ДНК волка, а d — двоичное слово, кодирующее ДНК собаки, то искомая величина есть $KS(d|w)$.

2.3 Колмогоровская сложность и шенноновская энтропия

Пусть случайная величина ω принимает n значений с вероятностями $p_1, p_2, ..., p_n$. Тогда ее шенноновская энтропия определяется формулой

$$H(\omega) = \sum p_i(-\log_2 p_i)$$

Говорят, что исход, имеющий вероятность p_i, несет в себе $(-\log_2 p_i)$ битов информации; тогда $H(\omega)$ — среднее количество информации в исходе случайной величины.

Оценим количество информации в ДНК. Допустим, количество всех существовавших ДНК всех организмов не превосходит 21000, все элементы этого множества равновероятны, тогда получится абсурдно малое число — меньше тысячи битов.

Теорема (о словах ограниченной сложности). Пусть n — произвольное число. Тогда существует менее 2^n слов x, для которых $KS(x) < n$.

Доказательство. Пусть D — оптимальный способ описания, фиксированный при определении колмогоровской сложности. Тогда все слова вида $D(y)$ при $l(y) < n$ (и только они) имеют сложность менее n. А таких слов не больше, чем различных слов y, имеющих длину меньше n, которых имеется

$$1 + 2 + 4 + 8 + ... + 2^{n-1} = 2^n - 1$$

штук (слов длины k ровно 2^k).

Отсюда можно заключить, что доля слов сложности меньше $n - c$ среди всех слов длины n меньше $\frac{2^{n-c}}{2^n} = 2^{-c}$. Например, доля слов сложности менее 90 среди всех слов длины 100 меньше 2^{-10}.

2.4 Сложность и случайность

Как мы уже видели, большинство слов несжимаемы или почти несжимаемы: скорее всего, случайно взятое слово данной длины окажется несжимаемым. Рассмотрим эксперимент: бросим монету, скажем, 80000 раз и сделаем из результатов бросаний файл длиной 10000 байтов. Можно утверждать, что ни один существовавший до момента бросания архиватор не сумеет сжать этот файл более чем на 10 байтов (вероятность этого меньше 2^{-80} для каждого архиватора, коих не так уж много)

Естественно считать случайными несжимаемые объекты: случайность есть отсутствие закономерностей, которые позволяют сжать объект. Конечно, нельзя провести четкой границы между случайными и неслучайными объектами. Например, какие именно из восьми слов длины 3 случайны, а какие нет? Определим дефект случайности слова x как разность $l(x) - KS(x)$. Тогда теорема о словах ограниченной сложности гласит, что для случайно взятого числа длины n вероятность иметь дефект больше d оценивается сверху числом $\frac{1}{2}d$.

Теорема. Пусть k — произвольная вычислимая функция из множества двоичных слов в N. Если k является нижней оценкой для колмогоровской сложности (то есть $k(x) \leq KS(x)$ для тех x, для которых $k(x)$ определена), то k ограничена.

Доказательство. По условию k — нижняя оценка колмогоровской сложности. Рассмотрим функцию $B(N)$, аргумент которой — натуральное число. $B(N)$ будем вычислять следующим алгоритмом: "развернуть параллельно вычисления $k(0), k(1), k(2), ...$ и проводить их до тех пор, пока не обнаружится некоторое x, для которого $k(x) > N$. Первое из таких x и будет результатом". Если k ограничена, то теорема доказана. В противном случае функция B определена для всех N, и $k(B(N)) > N$ по построению. По предположению k является нижней оценкой сложности, так что и $KS(B(N)) > N$. $B(N)$ эффективно получается по двоичной записи числа N, поэтому

$$KS(B(N)) \leq KS(bin(N)) + O(1) \leq l(bin(N)) + O(1) \leq \log_2 N + O(1)$$

Получаем, что $N < KS(B(N)) \leq \log_2 N + O(1)$, что при больших N приводит к противоречию.

3 Применение

Средства анализа колмогоровской сложности позволяют сравнительно несложно доказывать некоторые теоремы в области математической логики (к примеру, теорему Геделя о неполноте). Существует простая связь между колмогоровской сложностью и энтропией, позволяющая решать вероятностные задачи по аналогии.