О РАНГОВЫХ РАСПРЕДЕЛЕНИЯХ В КЛАССИФИКАЦИИ

Вадим Дунаев

Статья, впервые опубликованная в журнале "Научно-техническая информация. Серия 2, 1984г."

Поводом, но не главной целью, написания этой статьи явилось замечание Ю.А.Шрейдера (замечательного ученого и, по случаю, рецензента моей диссертации), что распределения типа Ципфа-Мандельброта не порождаются чисто стохастическими моделями. Нестохастическую природу этих распределений я тогда интуитивно остро чувствовал и понимал все то, о чем говорит мэтр, но сознательно мне хотелось найти простой стохастический генератор распределений данного типа, хотя бы из чувства противоречия. И вправду, я желал быть опровергнутым, но делал все, чтобы этого не призошло. Именно в этот момент появилась статья Гуссейна-Заде о том, как можно получить распределение типа Ципфа-Мандельброта посредством некоторого замысловатого стохастического источника. Данная статья была переполнена техническими подробностями, из-за чего я благодарен автору, искренне занимавшемуся исследованиями, чтобы приблизиться к истине, насколько это возможно. Эта статья мне очень понравилась своей изобретательностью, однако она была чересчур изысканной (как я тогда подумал, а теперь — сомневаюсь). Красивый результат, как мне тогда казалось, желательно было достичь короче, а, стало быть, внятнее и ближе к истине. Последнее, пожалуй, главное. Тогда я не нашутку задумался и вот, появилась данная статья.
Вспоминаю, вскоре после опубликования данной статьи в НТИ я получил письмо из Бельгии с просьбой выслать авторский препринт. Забавно, что меня довольно быстро нашли по незатейливому адресу "Russia, Red Army".
Я благодарен уважаемым Ю. Шрейдеру и С.Гусейну-Заде, благодаря которым я обратил свое внимание к одной из интереснейшей тем.

Данное введение к своей статье я написал спустя четверть века, т.е. в конце апреля 2009г.

§ 1. ВВЕДЕНИЕ

Ранговые распределения возникают, как известно, при изучении количественных характеристик текстов различной природы (литературных, музыкальных и живописных произведений, программ для ЭВМ), классификационных схем и других конструкций. Исследование механизмов, приводящих к построению объектов, характеризующихся ранговыми распределениями, является одним из направлений анализа интеллектуальной деятельности человека и находит приложения в автоматизированных информационных системах.

Среди ранговых распределений особое место занимает распределение, описываемое формулой Ципфа или ее обобщением — формулой Мандельброта. Вместе с тем существует достаточно много ситуаций, не согласующихся с указанными формулами (законами). Это и стимулировало исследования в двух основных направлениях: выбор зависимостей, более точно аппроксимирующих эмпирические данные, с одной стороны, и теоретическое обоснование закона Ципфа—Мандельброта, — с другой.

В рамках первого направления обычно модифицировалась формула Мандельброта введением дополнительных коэффициентов. Однако, как отмечалось, например, в [1], интерпретация результата в этом случае может только ухудшиться. В этой же работе была предложена новая зависимость, выведенная из довольно простых предположений. Выявленные в [1] механизмы, приводящие к

зависимости, которая существенно отличается от зависимости Ципфа—Мандельброта и хорошо описывает распределения реальных совокупностей объектов (например, ключевых слов, букв алфавита и других), позволяет отнести ее к ряду работ, связанных с теоретическим обоснованием ранговых распределений. Работа [1] и явилась, в некотором смысле, поводом к написанию настоящей статьи.

Ко второму направлению относится, прежде всего, результат, полученный Мандельбротом о связи зависимости Ципфа с оптимальным кодированием (например, [2]), вывод этой зависимости Ю. А. Шрейдером и М. В. Араповым на основе общесистемного принципа минимума симметрии ([3 и библиогр.]), работы Ю. К. Орлова (например, [4 и библиогр.]), обоснование Ю. А. Шрейдером закона Ципфа на основе понятия сложности [5] и другие. Однако, как уже отмечалось в [1], существующие обоснования используют довольно трудно интерпретируемые предположения. Поэтому представляет интерес поиск более простых предположений и рассуждений, приводящих к распределениям Ципфа—Мандельброта. В настоящей статье предлагаются несколько таких схем. Причем схемы, приведенные в § 2 настоящей статьи, являются по существу дальнейшим развитием идей, сформулированных в [5] еще тогда, когда понятие алгоритмической сложности только начинало входить в науку.

Прежде всего напомним основные сведения о законе Ципфа—Мандельброта. Невозрастающую последовательность

$$p_1,\ p_2, \ldots, p_k\ (\textstyle\sum^{k}_{i=1} p_i = 1)$$

частот употребления слов из словаря объема k в некотором тексте (выборке) называют ранговым распределением для данного текста. При этом номер i слова в словаре, упорядоченном по невозрастанию частоты употребления, называют рангом этого слова. Если F_i — количество употреблений слова ранга i (i-го слова), N — общее количество словоупотреблений в тексте, то $pi = Fil N$. Аналогичные характеристики используются при анализе разбиений классификационных универсумов на непересекающиеся классы: F_i — объем i-го класса, N — объем классификационного универсума, k — количество классов разбиения, p_i — относительный объем i-го класса. Текст (разбиение) удовлетворяет закону Ципфа, если его ранговое распределение описывается зависимостью

$$p_i = A/i.\ ,\ \textstyle\sum^{k}_{i=1} p_i = 1,\ i = 1,\ldots,k \qquad \textbf{(1)}$$

Условие нормировки частот однозначно определяет величину A:

$$A = p_1 \approx 1/\ln(k) \qquad \textbf{(2)}$$

Из выражения (2) и равенства $pi=F_i/N$ следует $N=Fi\ \ln(k)$.

В распределении Ципфа величины pi и k жестко взаимосвязаны. При анализе реальных текстов на предмет проверки соответствия закону Ципфа обычно задаются какой-нибудь одной из этих величин, наблюденной в данном тексте, а другую вычисляют, используя выражение (2). При этом распределение Ципфа будет зависеть от того, какая величина, p_i или k, определена по реальному тексту. По этой же причине может быть различным расхождение реального распределения и распределения Ципфа, зависящее от того, какое именно распределение Ципфа было выбрано в качестве эталонного при сравнении.

Выбрать эталонное распределение с учетом одновременно двух наблюденных величин, p_i и k позволяет формула Мандельброта, включающая формулу (1) как частный случай:

$$Pi= A/(i+B),\ \textstyle\sum^{k}_{i=1} p_i = 1,\ i = 1,\ldots,k \qquad \textbf{(3)}$$

Коэффициенты A и B в формуле (3) могут быть выражены через p_i и k с помощью следующих двух равенств:

$$A/(1+B) = p_1,\ \textstyle\sum^{k}_{i=1} A/(i+B) \approx A\ln((k+B)/(1+B)) = 1$$

Говоря о распределении Ципфа или Мандельброта, нередко принимают во внимание условие, согласно которому наименее употребительное слово встречается в тексте один раз (минимальный по объему класс содержит один элемент): $F_k=1$, $p_k=1/N$. При выполнении этого условия для текста, удовлетворяющего закону Ципфа, $F_1 = k$ и

$$A=p_1=1/\ln(F_1)=1/\ln(k),\ \ N=k\ln(k)=F_1\ln(F_1), \qquad \textbf{(4)}$$

а для текста, удовлетворяющего закону Мандельброта, —

$$A=1/\ln(F_1),\ B=(k-1)/(F_1-1)-1,\ N=(k-1)/(F_1-1)F_1\ln(F_1) \qquad \textbf{(5)}$$

Замечание. Здесь рассматривается частный случай формулы Мандельброта $p_i = A/(i+B)^\gamma$ при $\gamma = 1$

В дальнейшем, когда различия между распределениями Ципфа и Мандельброта для нас будут несущественны, мы будем говорить «распределение (закон) типа Ципфа».

Замечание. Для каждого конкретного текста (разбиения) можно подобрать наиболее близкое к нему (например, в среднеквадратическом смысле), наилучшее распределение типа Ципфа. При этом параметры p_i и k этого распределения могут и не совпадать с соответствующими реально наблюденными величинами. Тогда расхождения между реальным и наиболее близким к нему распределением типа Ципфа можно объяснить различиями в этих параметрах.

§ 2. К ОБОСНОВАНИЮ ЗАКОНА ТИПА ЦИПФА

Реальные тексты и разбиения, однако, довольно редко удовлетворяют закону Ципфа в точности. Вместе с тем было замечено, что тексты и классификационные схемы, хорошо согласованные с этим законом, соответствуют нашему интуитивному представлению о сбалансированности, целостности, системности. В то же время случайные выборки (случайным образом отобранные части целостного текста или, наоборот, конгломераты таких текстов) значительно хуже описываются зависимостью типа Ципфа [3,4]. Это обстоятельство и мотивировало рассмотрение степени соответствия текста (разбиения) закону Ципфа как меры его целостности, связности, хорошей организованности и т. п. Кроме того, оно вызвало вопрос: в каком же смысле хорош «идеальный» текст, в точности соответствующий закону типа Ципфа, и каков механизм действий (или схема рассуждений), приводящий к построению таких текстов?

Как показал Б.Мандельброт, появление распределений типа Ципфа связано с оптимальным кодированием. Однако на практике, особенно при написании больших текстов (описании через признаки классов разбиения больших классификационных универсумов), объект кодирования (объект нашей мысли) формируется или обозревается нами не целиком, а по частям, рассматриваемым последовательно во времени. При этом кодирование объекта мысли происходит уже тогда, когда он еще не сформировался окончательно. Не исключено, что издержками такого способа кодирования являются, например, вхождения в один и тот же текст одинаковых слов с различными значениями и, наоборот, — различных слов с одинаковыми значениями. Очевидно, что такое кодирование, вообще говоря, оптимальным быть не может. Однако можно предположить, что оно может быть «локально оптимальным» (оптимальным на каждом шаге кодирования отдельных частей объекта) и, более того, при определенных условиях оно может приводить к результату, «почти» такому же, как и оптимальное кодирование («глобально оптимальное»). Ниже мы рассмотрим два способа кодирования, отвечающие этой гипотезе и уточняющие ее, которые приводят к распределению типа Ципфа. Оба способа мы рассмотрим применительно к задаче описания классов разбиения через признаки.

2.1. Последовательное описание классов через признаки. Пусть необходимо описать k классов некоторого разбиения классификационного универсума через m-значные признаки (число $m\geq 2$, как мы увидим, не играет принципиальной роли). Описать класс через признаки означает сопоставить ему кортеж значений некоторых признаков, которыми характеризуются все объекты данного класса и только они. Мы будем рассматривать тот случай, когда каждому классу можно сопоставить единственный кортеж значений признаков. Вопрос о том, каким образом и какие признаки следует использовать,

здесь не рассматривается: нас интересует только количественная сторона дела. Объем N классификационного универсума и объемы F_i классов заранее фиксировать не будем.

Предположим, что классификатор описывает классы последовательно: сначала какой-нибудь один класс, затем другой и т. д. При этом необходимое количество признаков, используемых для описания класса, определится в соответствии с логикой оптимального m-ичного поиска (выбора) сначала одного класса из множества k классов, затем одного класса из оставшихся неописанными k—1 классов и т. д.—до тех пор, пока не будут исчерпаны все k классов данного разбиения.

Для выбора одного элемента из множества k элементов требуется, как известно, не более ближайшего к $\log_m(k)$ не меньшего целого числа m-значных признаков, причем максимальное число признаков не может быть уменьшено. В дальнейшем мы будем использовать приближенную оценку $\log_m(k)$ для числа признаков. Итак, допустим, что выбранный класс описан через $\log_m(k+B)$ признаков, где $B \geq 0$ характеризует некоторый запас признаков.

После описания i-го класса следует описать (i+1)-й класс, выбрав его из оставшихся k—1 классов и подобрав $\log_m(k+B-i+1)$ подходящих признаков, и т. д. При этом классы окажутся упорядоченными по неувеличению числа признаков. Изменим нумерацию классов на противоположную. Тогда число признаков i-го класса будет равно с принятой здесь точностью $\log_m(i+B)$.

Предположим, далее, что при описании классов используются значения признаков, взятых из множества уже использованных при описания k-го класса (в новой нумерации). Иначе говоря, множество всех признаков, выбранных для описания всех классов, совпадает с множеством признаков, через значения которых описан *k-й* класс. В этом случае объекты k-го класса неразличимы по выбранным признакам и поэтому будем считать, что k-й класс содержит единственный объект. Тогда объем 1-го класса

$$F_i = m^{lmax-l(i)},$$

где $l\ (i) = \log_m(i + B)$ — число признаков, использованных при описании i-го класса; $l_{max} = l\ (k)$.

В результате очевидных преобразований получаем:

$$F_i = (k+B)/(i+B),\ B=(k-1)/(F_1-1)-1;$$

$$N = \sum_{i=1}^{k} F_i \approx (k+B)\ln((k+B)/(1+B)) = (k-1)/(F_1-1)\ln(F_1));$$

$$p_i = Fi/N = 1/((i+B)\ln((k+B)/(1+B))) = 1/((i+B)\ln(F_1))$$

Нетрудно заметить (ср. с выражениями (3) и (5)), что рассмотренная схема последовательного описания классов приводит к распределению Мандельброта. Если при этом ограничиться минимальным количеством признаков (B = 0), то получится распределение Ципфа.

Замечание. При оптимальном описании классов, относительные объемы которых распределены по закону Мандельброта, т. е. при описании, обеспечивающем минимум среднего количества признаков в расчете на один класс, количество признаков для i-го класса приближенно равно

$$-\log_m Pi = \log_m(i+B) + \log_m \ln F_1.$$

При рассмотренном выше способе описания количество признаков для i-го класса равно $\log_m(i+B)$, т. е. несколько меньше.

Однако асимптотически обе эти величины эквивалентны, т. е. при $k \to \infty$ и $i \to k$ их отношение стремится к 1. В этом смысле можно понимать «почти» оптимальность алгоритма последовательного описания классов через признаки.

2.2. Кодирование классов натуральными числами. Пусть $H(x)$— Сложность описания объекта x из классификационного универсума U, $H(x/y_i)$— сложность описания объекта x, если известно, что он принадлежит классу y_i заданного разбиения U на непересекающиеся классы (можно принять, что $H(x) = H(x/U)$, $I(x:y_i)$— количество информации в объекте x о классе y_i. Как известно из алгоритмической теории информации (например, [6]), между указанными величинами имеет место следующее соотношение:

$$H(x/y_i)+I(x:y_i) = H(x). \qquad \textbf{(6)}$$

Дальнейшие рассуждения опираются на способ определения данных величин. Определим сложность описания объекта x как минимальное количество m-значных признаков, по которым он может быть выделен из известного множества:

$$H(x) = \log_m|U| = \log_m N,$$

$$H(x|y_i) = \log_{m|}|y_i| = \log_m F_i.$$

Иначе говоря, $H(x/y_i)$ есть минимальное число шагов m-ичного алгоритма поиска, останавливающегося при выделении объекта из известного класса y_i. Далее, будем считать, что информация в объекте x о классе y_i заключается в указании номера i класса y_i, которому принадлежит x. Количество этой информации определим через длину записи числа i в некотором m-ичном алфавите. Как известно, длина записи натурального числа i в m-ичном алфавите приближенно равна целой части $\log_m(i+1)$. Таким образом, $I(x : yi) \approx \log_m(i+1)$. Однако мы примем, что $I(x : y_i) \geq \log_m \alpha_i$, где $\alpha \geq 1$—некоторый коэффициент. Тогда с учетом приведенных выше определений равенство (6) можно переписать в виде

$$\log_m(\alpha_i) = \log_m (N/F_i) = \log_m (1/p_i), \qquad \textbf{(7)}$$

откуда получаем распределение Ципфа $p_i = 1/(\alpha_i)$, где $\alpha = \ln(k)$ получается из условия нормировки.

Нетрудно заметить, что формула Ципфа могла быть получена непосредственно из предположения, что кодирование классов натуральными числами, представленными в некотором m-ичном алфавите, «почти» оптимально в смысле минимума средней длины кодового слова, т. е. из предположения, соответствующего равенству (7).

2.3. Обобщение. Последовательный перебор (2.1) и нумерация (2.2) объектов являются операциями, которые неизбежно, как нам представляется, используются в классификационной деятельности. Применение этих операций индуцирует некоторые разбиение классификационного универсума, классы которого могут быть упорядочены по численности. Если, к тому же, операции рациональны (близки к оптимальным в смысле поиска или кодирования), то получающиеся ранговые распределения численности классов могут быть описаны с некоторым приближением формулой типа Ципфа.

Разумеется, это — не единственная система предположений, которая может быть положена в основание закономерности Ципфа. Как будет показано в последующих разделах, закономерность Ципфа может быть получена, если в схему рассуждений привнести элементы случайности.

§ 3. СРЕДНЕСТАТИСТИЧЕСКИЕ РАНГОВЫЕ РАСПРЕДЕЛЕНИЯ

Пусть объемы F классов некоторого разбиения являются независимыми случайными величинами с одинаковыми непрерывными функциями распределения вероятностей $W(F)$.

Тогда [7,с. 512—515] распределение вероятностей i-го по величине объема ($i=1,\ldots, k$) асимптотически (при $k\to\infty$) подчиняется нормальному закону с математическим ожиданием, приближенно равным

$$\mathbf{F}_i = W^{-1}((k-i+1)/(k+1)) \qquad \textbf{(8)}$$

где W^{-1}—функция, обратная W.

Если $W(F) = 1 — e^{-k/NF}$, то выражение (8) можно переписать в виде:

$$1-e^{-(k/N)\mathbf{F}i} = (k-i+1)/(k+1),$$

откуда следует ранговое распределение, предложенное Гусейном-Заде [1]:

$$\mathbf{p_i} = \mathbf{F_i}/N = (1/k)ln((k+1)/i) \qquad \textbf{(9)}$$

Как отмечалось в [1], формула (9) хорошо описывает распределение ключевых слов, букв алфавита, первых букв фамилий. Важную роль в выводе среднестатистического рангового распределения (9) играет то, что объем F распределен по экспоненциальному закону. Из статистической теории информации известно, что такое распределение имеет максимальную энтропию при ограничениях на число классов и объем классификационного универсума или, наоборот, минимизирует объем классификационного универсума при ограничениях на число классов и энтропию.

Та же схема рассуждений, но при $W(F)=k- 1 /F$ приводит к $\mathbf{F}_i = k/i$, откуда с учетом равенства $\sum^{k}_{i=1} \mathbf{F}_i = \mathbf{N} \approx kln(k)$ получаем формулу Ципфа

$$\mathbf{P}_i = \mathbf{F}_i/\mathbf{N} = 1/(iln(k))$$

Покажем теперь, каким образом может быть получено распределение вероятностей $W(F) = 1- 1/F$.

Хотя закон Ципфа является скорее законом отдельных связанных текстов, чем языка вообще [4], он, тем не менее, в некотором приближении выполняется и на больших совокупностях текстов, представляющих какую-то часть языка. Так, частоты первых 30 наиболее употребительных слов английской газетной лексики (выборка $2\cdot10^5$ словоупотреблений со словарем объемом 12 588 слов [8]) в десятки раз лучше описывается формулой Ципфа, чем зависимостью Гусейна-Заде (9). Аналогично можно предположить, что относительные объемы классов «естественного» разбиения некоторого достаточно разнородного классификационного универсума, подобно частотам слов в языке, также подчинены закону Ципфа. Однако в дальнейших рассуждениях представляется более надежным использовать языковую интерпретацию.

Итак, допустим, что с достаточной для нас точностью частоты слов языка со словарем объема k описываются формулой Ципфа. Тогда количество употреблений i-го слова $F_i=k/i$. Учет требования, согласно которому F_i должно быть натуральным числом, приводит к тому, что некоторые различные слова должны иметь одинаковые количества употреблений. Пусть для некоторого $i = 1, \ldots, k$ $F_i=F$, где F — натуральное число. Тогда число $k\{F)$ различных слов с количеством употреблений, равным F, можно приближенно определить из условия $F_{i-k(F)} = F+1$: $k(F) = k/(F(F+l))$. Последнее выражение является достаточно точным при небольших F, пока погрешность округления $k(F)$ до ближайшего целого числа относительно мала.

Пусть слово из словаря выбирается случайным образом (с вероятностью $1/k$). Тогда вероятность того, что это слово будет употреблено F раз,

$$w(F) = k(F)/k = 1/(F(F+1)),$$

а функция распределения

$$\mathrm{W(F)} = \sum^{F}_{i=1} w(x) \approx \int^{F}_{1} w(F)dx = ln3 - ln(1+1/F) \approx 1-1/F \qquad (10)$$

Дискретное распределение вероятностей (10) можно аппроксимировать непрерывной функцией такого же вида, положив, что F принимает значения из интервала $[1, \infty)$ действительных чисел.

Скажем, что употребление в тексте некоторых слов из словаря языка с о г л а с о в а н о с языком, если количество их употребления в тексте распределено по закону (10). Тогда случайный и

независимый выбор слов, употребляемых для написания текста в согласии с языком, приводит к таким ранговым распределениям, что математическое ожидание частоты ранга i удовлетворяет закону Ципфа.

Как уже отмечалось, близость реального распределения к распределению типа Ципфа является признаком системности, целостности соответствующего текста. Однако кроме этого критерия можно использовать еще и степень удаленности реального распределения от распределения Гусейна-Заде (9), поскольку вывод последнего существенно использует факторы случайности. Это выражается, в частности, в появлении экспоненциальной функции распределения вероятностей $W(F)$, которое обычно связано с марковскими процессами.

Например, формула (9) хорошо описывает ранговое распределение частот букв алфавита (формула Ципфа — плохо), плохо — распределение первых частот слов английской газетной лексики (формула Ципфа — неплохо) и лучше, чем формула Ципфа, описывает распределение словосочетаний английской газетной лексики [8]. Возникает подозрение, что формула (9) лучше соответствует распределениям таких объектов, связь между которыми в тексте слабее, а формула типа Ципфа, наоборот, лучше описывает распределения взаимобусловленных объектов. Однако это — только гипотеза, нуждающаяся в более тонком экспериментальном подтверждении.

Замечстие. Отметим еще один способ, приводящий к распределению типа Ципфа. Пусть дан текст с ранговым распределением частот k различных слов, описываемым формулой Ципфа. Разобьем множество (объема $N = kln(k)$) всех вхождений слов в данный текст на непересекающиеся классы следующим образом: вхождения слов x и y принадлежат i-му классу ($i=1, 2,\ldots$), если слова x и y входят в текст по i раз каждое. Поскольку число различных слов, использованных в тексте i раз, равно $k(i) = k/(i(i+1))$, то объем i-го класса рассматриваемого разбиения равен $ik(i)$, а относительный объем — $ik(i)/N=1/((i+1)/ln(k))$, т. е. описывается формулой, близкой к формуле Ципфа.

§ 4. О ДУАЛЬНЫХ РАСПРЕДЕЛЕНИЯХ, ВОЗНИКАЮЩИХ ПРИ СИНТЕЗЕ АЛГОРИТМОВ КЛАССИФИЦИРОВАНИЯ

В этом разделе мы рассмотрим ситуацию, когда классификатор описывает классы некоторого разбиения классификационного универсума через признаки и сопоставляет описанным классам их имена, представленные в некотором коде. Результатом такой деятельности может быть дерево (древовидный алгоритм) классифицирования, неконцевым вершинам которого соответствуют признаки, дугам — значения соответствующих признаков, а концевым вершинам — имена классов в некотором коде. В основу дальнейших рассуждений мы положим гипотезу, согласно которой при синтезе древовидных алгоритмов классифицирования объектов из больших универсумов на большое число непересекающихся классов классификатор стремится, осознанно или нет, к тому, чтобы выполнялось (в некотором приближении) соотношение (6) со следующими оговорками.

Под объектами классификационного универсума численностью N будем понимать всевозможные кортежи значений $log_m N$ некоторых m-значных (m≥2) признаков, так что $H(x) = log_m N$. Величина $I(x{:}y_i)$ равна длине представления имени (номера) i-го класса в некотором коде. Через s_i обозначим число m-значных признаков, кортеж значений которых сопоставлен i-му классу. Предположим, что s_i равно $H(x|y_i)$ с точностью до адитивной неотрицательной константы β,

которая определяется неравенством Крафта:

$$\sum_{i=1}^{k} m^{s_i} \leq 1,$$

k — число классов (необходимое и достаточное условие существования m-ичного однозначно декодируемого кода с длинами кодовых слов s_i [2]). Тогда выражение (6) перепишется в виде:

$$s_i + I(x : y_i) = log_m(N) + \beta \qquad (11)$$

Предположим, что номера классов кодируются оптимальным n-ичным кодом (кодом Хаффмена [2]). Тогда $I(x : y_i) \approx -log_n q_i$, где q_i —вероятность того, что предъявленный для классифицирования объект

принадлежит i-му классу; будем называть q_i вероятностью i-го класса. Поскольку объем F_i i-го класса, которому сопоставлен кортеж значений признаков длины s_i,

$$F_i = m^{logm(N-si)},$$

равенство (11) можно переписать в виде:

$$F_i q_i^{log_n m} = m^{-\beta}$$

В частном случае бинарных признаков и кода ($m= n=2$) это выражение упрощается: $F_i q_i = 2^{-\beta}$. Когда неравенство Крафта обращается в равенство, выражение это можно записать в следующем виде:

$$p_i q_i = 1/(\sum^{k}_{i=1}(1/q_i)) \qquad (12)$$

где p_i — F_i/N — относительный объем i-го класса. Если классы упорядочить по неувеличению их вероятности ($q_1 \geq q_2 \geq ... \geq q_k$), то они окажутся упорядоченными по неуменьшению их относительных объемов ($p_1 \leq p_2 \leq ... \leq p_k$). В этом смысле распределения q_i и p_iдуальны. Дуальность этих распределений была эмпирически подмечена автором при разработке древовидных алгоритмов классифицирования для оценки состояния сложных технических устройств, когда объекты представлялись в виде кортежей значений десятков измеряемых параметров (состав параметров был фиксирован), а количество классов заранее не фиксировалось, но в конечном итоге оказывалось больше 100. Эти числа при синтезе алгоритмов классифицирования вполне можно считать большими.

Возможно, что появление дуальных распределений при решении задач классификации связано с особенностями организации человеческой памяти, с одной стороны, и оптимизацией кодирования при коммуникациях,— с другой. Так, учитывая, что $si \approx -log_m p_i$, представляется рациональным распределение памяти, согласно которому наиболее вероятные классы описываются через относительно большое число признаков (отводится много места для описания таких классов), и, наоборот, — описания маловероятных классов занимают относительно мало места (классы с нулевой вероятностью вообще не описываются). Кроме того, оптимальное распределение времени на передачу результатов классифицирования должно быть противоположным: наиболее вероятные классы должны кодироваться относительно короткими именами, а маловероятные классы — длинными именами. Так, о тех объектах, с которыми часто приходится иметь дело, мы обычно многое можем рассказать (припомнить много признаков), но при общении говорим о них кратко. И наоборот, объекты, с которыми ты сталкиваемся редко, оставляют небольшой след в нашей памяти, но при попытке рассказать о них нередко приходится прибегать к многословию.

Рассмотрим еще одну интерпретацию дуальных распределений. Пусть классификационный универсум состоит из видов, которые, в свою очередь, состоят из индивидов (особей); F_i — количество видов i-го класса разбиения классификационного универсума на k классов; r_{ij} — количество индивидов j-го вида i-го класса ($j = 1, \ldots, F_i$). Тогда $F^+_i = \sum^{Fi}_{i=1} r_{ij}$ — количество индивидов i-го класса, а $N^+ = \sum^{k}_{i=1} F^+_i$ — количество всех индивидов, принадлежащих видам классификационного универсума. Полагая появление индивидов случайным и независимым с равномерным законом распределения, вероятность i-го класса можно представить в виде $q_i = F^+_i/N^+$. Тогда при выполнении равенства $p_i q_i = const$, $i=1,...,k$ получается, что классы с большим числом видов относительно малопредставительны и, наоборот, классам с небольшим числом видов соответствует относительно много индивидов. Наблюдается ли подобное соотношение для некоторых известных крупных классификационных схем, получивших репутацию естественных, автору не известно, и поэтому данную интерпретацию можно рассматривать не более как гипотезу.

Оказывается, что распределение Ципфа дуально к ранговому распределению, которое можно получить на основе наиболее простой статистической модели. Эта модель аналогична рассмотренной в § 3 с той лишь особенностью, что объемы классов F подчинены равномерному закону с функцией распределения вероятностей $W\{F)=F/k, \ 0<F\leq k$. Напомним, что все приведенные ниже оценки имеют

асимптотический ($k\to\infty$) характер. Из выражения (8) следует, что математическое ожидание объема i-го по величине класса

$$\mathbf{F}_i = k(k-i+1)/(k+1),\ i = 1, \ldots, k.$$

Перенумеруем классы в порядке неуменьшения $\mathbf{F}_i$ тогда

$$\mathbf{F}_i = ik/(k+1).$$

Так как $N = \sum_{i=1}^{k}\mathbf{F}_i = k^2/2$, то $\mathbf{p}_i = \mathbf{F}_i/\mathbf{N} = 2i/(k(k+1))$.

Из (12) следует, что распределение q_i, дуальное к p_i, описывается формулой Ципфа:

$$q_i = 1/(\mathbf{p}_i\sum_{i=1}^{k}(1/\mathbf{p}_i)) = 1/(i\sum_{i=1}^{k}(1/i)) \approx 1/(i\ln(k))$$

ЛИТЕРАТУРА

1. Гусейн-Заде С. М. О встречаемости ключевых слов и о других ранжированных рядах // НТИ. Сер, 2.— 1987. — № 1. —С. 28—32.

2. Брюллюэн Л. Наука и теория информации. — М.: Физматгиз, I960. —392 с.

3. Шрейдер Ю. А., Шаров А. А. Системы и моде ли.— М.: Радио и связь, 1982.— 152 с.

4. Орлов Ю. К. Невидимая гармония//Число и мысль. Вып. 3. —М;: Знание, 1980.— С. 70—106.

5. Шрейдер Ю. А. О возможности теоретического вывода статистических закономерностей текста // Проблемы передачи информации.— 1967. — Т. 3, вып. 1. —С. 57—63.

6. Успенский В. А., Семенов А. Л. Теория алго ритмов: Основные открытия и приложения. — М.5 Наука, 1987. —288 с.

7. Кокс Д., Хинкли Д. Теоретическая статистика.— М.: Мир, 1978..—560 с.

8. Алексеев П. М., Турыгина Л. А. Частотный англо-русский словарь-минимум газетной лексики. — М.: Воениздат, 1974.— 261 с.

К 80-летию Б. Мандельброта