

Оптимальная обработка вероятностных прогнозов двух экспертов*

Алексей Львович Семенов[†] Андрей Альбертович Мучник[‡]

*Вычислительный центр РАН, отделение кибернетики,
119991, Москва, ГСП-1*

Аннотация

Рассматривается прогнозирование в смысле Соломонова последовательностей в конечном алфавите A (inductive inference). Собственно прогнозами называются распределения вероятностей на A . Распределение p интерпретируется как прогноз для n -го члена последовательности y_1, y_2, \dots , если с точки зрения автора прогноза при известных значениях y_{n-1}, y_{n-2}, \dots условная вероятность события $y_n = a$ равна $p(a)$.

Неформально, цель состоит в построении следующей процедуры. Она получает на вход несколько последовательностей прогнозов для первых n членов некоторой последовательности y , а также значения первых $n - 1$ членов y . На выход даётся ещё один прогноз для n -го члена y . Предположим, что для каждого члена y прогноз из j -ой последовательности исходит от j -го эксперта. Тогда мы стремимся, чтобы последовательность прогнозов на выходе процедуры оказалась достаточно точной (в вероятностном смысле) с точки зрения каждого эксперта. Для измерения точности прогнозирования используется расстояние Хеллингера.

Для ситуации, когда экспертов двое, найдена оптимальная (относительно рассматриваемой меры качества) процедура прогнозирования. Показана неоптимальность двух классических процедур: «взвешенное среднее» и «максимальное правдоподобие».

1. Несколько слов о выборе функции расстояния

Если на конечном или счётном множестве заданы два распределения вероятностей p и q , то расстояние между ними мы будем измерять по Хеллингеру: $\rho(p, q) = \sum_i (\sqrt{p(i)} - \sqrt{q(i)})^2$. Есть и другие способы измерения расстояний. Назовём некоторые причины, по которым мы будем использовать именно расстояние Хеллингера.

*Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований, проекты № 04-01-00427, № 02-01-10904, и Совета поддержки научных школ при президенте РФ.

[†]e-mail: alsemenov@mtu-net.ru, fax: +(095)9156963.

[‡]e-mail: muchnik@lpcs.math.msu.su, fax: +(095)9156963.

Во-первых, в 1986 г. Вовк доказал следующее (см. [1]).

Пусть на пространстве бесконечных последовательностей в конечном алфавите $A = \{a_0, a_1, \dots\}$ (то есть на дереве ветвления $|A|$) даны две вычислимые меры P и Q , положительные на всех конусах. Мера P конуса над вершиной x будет обозначаться $P(x)$. Для $a \in A$ выражение $P(xa|x)$ будет обозначать $P(xa)/P(x)$. Если x фиксированно, то $P(xa|x)$ задаёт распределение вероятностей на A . Начало длины n бесконечной последовательности y обозначим через $y_{1:n}$.

Бесконечная последовательность y , случайная по Мартин-Лёфу относительно P , тогда и только тогда случайна по Мартин-Лёфу относительно Q , когда

$$\sum_n \rho(\lambda a.P(y_{1:n}a|y_{1:n}), \lambda a.Q(y_{1:n}a|y_{1:n})) < \infty$$

(в дальнейшем связи $\lambda a.$ будут опускаться).

Получить аналогичный критерий в терминах других расстояний не удаётся.

Во-вторых, функция ρ симметрична и ограничена (в отличие, например, от часто используемых в математической статистике расстояния Кульбака и расстояния χ^2). Очевидно, всегда $0 \leq \rho \leq 2$ и $\rho(p, q) = 0 \Leftrightarrow p = q$. Отметим важную особенность функции ρ . Если $\forall i (p(i) = 0 \rightarrow q(i) = 0) \wedge \forall i (p(i) \neq 0 \rightarrow |p(i) - q(i)| \leq \varepsilon \ll p(i))$, то $\rho(p, q) = O(\varepsilon^2)$. Если же $\exists i (p(i) = 0 \wedge q(i) = \varepsilon > 0)$, то $\rho(p, q) > \varepsilon$. Это отражает тот факт, что эффект замены нулевой вероятности на ненулевую бывает гораздо больше, чем эффект небольшого изменения положительной вероятности. Возможно, именно поэтому неудобно использовать расстояния

$$L_1(p, q) = \sum_i |p(i) - q(i)|;$$

$$L_2(p, q) = \sqrt{\sum_i (p(i) - q(i))^2};$$

$$L_\infty(p, q) = \max_i |p(i) - q(i)|.$$

С другой стороны, в отличие от L_1, L_2, L_∞ , функция ρ не является нормой: $\rho(p, q)$ бывает строго больше $\rho(p, r) + \rho(q, r)$ (например, для $p = (1, 0), q = (0, 1), r = (\frac{1}{2}, \frac{1}{2})$). Неравенство, близкое к неравенству треугольника, для ρ всё-таки есть: $\rho(p, q) \leq 2(\rho(p, r) + \rho(q, r))$.

2. Основные определения

Буквы алфавита A будем интерпретировать как некоторые события, n -я буква последовательности в алфавите A интерпретируется как событие, произошедшее в момент времени n . *Экспертом* назовём функцию, которая каждой конечной последовательности x букв из A сопоставляет распределение на A , обозначаемое $P(xa|x)$ (при условии, что в предыдущие моменты времени произошли события из x , вероятность того, что в следующий момент произойдёт событие a , с точки зрения эксперта равна $P(xa|x)$). Наша цель — по набору P_1, \dots, P_J экспертов построить ещё одного эксперта Q , предсказания которого были бы достаточно точны с точки зрения каждого

из имевшихся экспертов. *Суммарную ошибку* эксперта Q с точки зрения эксперта P на последовательности $\langle y_1, \dots, y_N \rangle$ определим как $\sum_{n=0}^N \rho(P(y_{1:n}a|y_{1:n}), Q(y_{1:n}a|y_{1:n}))$. Как это обычно для теории вероятностей, мы не запрещаем появление больших суммарных ошибок, но требуем, чтобы вероятность такого появления с точки зрения P была мала. То есть нас будет интересовать скорость убывания функции $\mathcal{F}_{P,Q}(s)$, равной мере P множества $\{y \mid \sum_{n=0}^{\infty} \rho(P(y_{1:n}a|y_{1:n}), Q(y_{1:n}a|y_{1:n})) > s\}$. (Очень малое изменение значений $P(xa|x)$, при котором все они становятся ненулевыми, не отражается на дальнейших результатах, но позволяет избежать вопросов, связанных с делением на нуль.)

Назовём *предсказателем* оператор, который, получив последовательность $\langle y_1, \dots, y_N \rangle$ и значения $P_j(y_{1:n}a|y_{1:n})$ для всех $j \in \{1, \dots, J\}$, $n \in \{0, \dots, N\}$, $a \in A$, строит распределение вероятностей на A , соответствующее $Q(y_{1:N}a|y_{1:N})$. (При $N = 0$ предсказатель получает на вход пустую последовательность.) Мы будем отдельно рассматривать *цельные* предсказатели, для которых $\forall N \exists j \forall a Q(y_{1:N}a|y_{1:N}) = P_j(y_{1:N}a|y_{1:N})$ и которые при выборе одного из j на шаге N не учитывают значений $P_j(y_{1:N}a|y_{1:N})$. Понятие цельного предсказателя особенно важно для тех приложений, когда нельзя частично довериться одному эксперту и частично другому.

Предположим, что в действительности последовательность y порождается в соответствии с распределением R . Тогда предсказания Q окажутся достаточно точными относительно R , если *хотя бы один* из экспертов P_j достаточно точен относительно R . (Под точностью понимается скорость убывания $\mathcal{F}_{R,Q}(s)$ и $\mathcal{F}_{R,P_j}(s)$.) Так как наши результаты будут касаться только функций $\mathcal{F}_{P_j,Q}(s)$, то никакие предположения о действительном распределении R (и даже о его наличии) в формулировках теорем нам не понадобятся.

В начале 1950-х годов Соломонов рассмотрел понятие предсказателя (inductive inference) и предложил класс предсказателей, называемых *взвешенным средним*. Припишем каждому эксперту P_j вес $w_j > 0$ так, что веса образуют распределение вероятностей на номерах экспертов. Предположим, что среди экспертов есть такой, который предсказывает «правильно», то есть его предсказания совпадают с объективными вероятностями R . Тогда вес w_j выражает степень нашей априорной уверенности в том, что указанным экспертом является P_j . После получения информации о первых n событиях ($\langle y_1, \dots, y_n \rangle = z$) наше доверие к j -му эксперту, естественно, может меняться. Оно вычисляется по классической формуле $w_j P_j(z) / \sum_i w_i P_i(z) = W_j(z)$. Предсказатель Q («взвешенное среднее») определяется так: $Q(za|z) = \sum_j W_j(z) P_j(za|z)$. Просто проверяется, что $Q(za|z) = \sum_j w_j P_j(za) / \sum_j w_j P_j(z)$ и что $W_j(za) = W_j(z) P_j(za|z) / \sum_i W_i(z) P_i(za|z)$. Таким образом, предсказателю Q для работы достаточно помнить апостериорные веса $W_j(z)$ и получать последнее предсказание каждого эксперта. Заметим, что предположение $P_j = R$ было нужно только для мотивировки. Формально, в определении это никак не используется.

С первого взгляда предсказатель «взвешенное среднее» кажется очень естественным. Теперь рассмотрим конечное множество экспертов и потребуем цельности предсказателя. Тогда наиболее естественным с первого взгляда кажется предсказатель *максимальное правдоподобие*. Для этого предсказателя $\forall a Q(y_{1:n}a|y_{1:n}) = P_j(y_{1:n}a|y_{1:n})$ для того (одного из тех) j , для которого $P_j(y_{1:n})$ принимает наибольшее значение. Для работы ему достаточно помнить числа $P_j(y_{1:n}) / \sum_i P_i(y_{1:n})$ и получать

последнее предсказание каждого эксперта.

3. Обзор результатов

Тот факт, что для предсказателя «взвешенное среднее» значения условных вероятностей $Q(y_{1:n+1}|y_{1:n})$ и $P_j(y_{1:n+1}|y_{1:n})$ сближаются при росте n для почти всех y в смысле P_j , был известен Соломонову с самого начала. По теореме Дуба для любых мер P и Q , заданных на дереве, отношение $Q(y_{1:n})/P(y_{1:n})$ имеет конечный предел при $n \rightarrow \infty$ для почти всех в смысле меры P последовательностей y . Поскольку для «взвешенного среднего» всюду выполнено $Q(x)/P_j(x) \geq w_j$, то почти всюду $\lim_{n \rightarrow \infty} Q(y_{1:n})/P_j(y_{1:n}) > 0$, и следовательно, выполняется равенство $\lim_{n \rightarrow \infty} Q(y_{1:n+1}|y_{1:n})/P_j(y_{1:n+1}|y_{1:n}) = 1$. Если условные вероятности $P(xa|x)$ всюду больше какой-нибудь положительной константы, то теорема Дуба имеет усиление: для каждой конечной последовательности z для почти всех относительно P бесконечных последовательностей y предел при $n \rightarrow \infty$ отношения $Q(y_{1:n}z)/P(y_{1:n}z)$ равен $\lim_{n \rightarrow \infty} Q(y_{1:n})/P(y_{1:n})$. Следовательно, при условии $\exists \delta > 0 \forall x, a P_j(xa|x) > \delta$ для почти всех относительно эксперта P_j бесконечных последовательностей y выполнено $\forall z, a \lim_{n \rightarrow \infty} Q(y_{1:n}za|y_{1:n}z)/P_j(y_{1:n}za|y_{1:n}z) = 1$, где Q — взвешенное среднее экспертов P_i . (Условие отделимости от нуля условных вероятностей ошибочно отсутствует в теореме 5.2.2 из [3, с. 332]. Существенность этого условия показывает следующий пример. Пусть P_1, P_2 — меры на бинарном дереве, и $\forall n P_1(0^n 1|0^n) = (n+2)^{-3}$, $P_2(0^n 1|0^n) = (n+2)^{-2}$. Обе меры положительны на последовательности 0^ω . Следовательно, апостериорные веса $W_1(0^n), W_2(0^n)$ имеют положительные пределы при $n \rightarrow \infty$. Тогда $Q(0^n 1|0^n) = \Omega((n+2)^{-2})$ и $\lim_{n \rightarrow \infty} Q(0^n 1|0^n)/P_1(0^n 1|0^n) = \infty$.)

Если мера Q построена предсказателем «максимальное правдоподобие», то для каждого j имеют место те же утверждения: для почти всех относительно P_j последовательностей y

$$\lim_{n \rightarrow \infty} Q(y_{1:n+1}|y_{1:n})/P_j(y_{1:n+1}|y_{1:n}) = 1;$$

если $\exists \delta > 0 \forall x, a P_j(xa|x) > \delta$, то для почти всех относительно P_j последовательностей y

$$\forall z, a \lim_{n \rightarrow \infty} Q(y_{1:n}za|y_{1:n}z)/P_j(y_{1:n}za|y_{1:n}z) = 1.$$

Действительно, в каждой вершине x условная вероятность $Q(xa|x)$ равна одной из условных вероятностей $P_i(xa|x)$. Поэтому предел $Q(y_{1:n+1}|y_{1:n})/P_j(y_{1:n+1}|y_{1:n})$ будет равен 1, если равны 1 пределы $P_i(y_{1:n+1}|y_{1:n})/P_j(y_{1:n+1}|y_{1:n})$ для тех i , для которых в бесконечно многих вершинах $y_{1:n}$ i -й эксперт наиболее правдоподобен. Для таких i $\exists^\infty n P_i(y_{1:n})/P_j(y_{1:n}) \geq 1$, и следовательно, предел $P_i(y_{1:n})/P_j(y_{1:n})$ может быть только положительным.

Теорема Дуба и основанные на ней рассуждения не дают никакой оценки на скорость сближения $Q(y_{1:n+1}|y_{1:n})$ и $P_j(y_{1:n+1}|y_{1:n})$. В 1978 г. Соломонов [2] получил такую оценку для предсказателя «взвешенное среднее». Он доказал, что для каждого j

$$\mathbf{E}_{P_j} \sum_{n=0}^{\infty} \rho'(P_j(y_{1:n}a|y_{1:n}), Q(y_{1:n}a|y_{1:n})) \leq -\ln w_j,$$

где ρ' — расстояние Кульбака ($\rho'(p, q) = \sum_a p(a) \ln(p(a)/q(a))$). Из этого следует, что мера P_j множества $\{y \mid \sum_{n=0}^{\infty} \rho'(P(y_{1:n}a|y_{1:n}), Q(y_{1:n}a|y_{1:n})) > s\}$ меньше $(-\ln w_j)/s$. Функция $(-\ln w_j)/s$ убывает слишком медленно. В статье [1] Вовк приводит конструкцию, из которой видно, что для расстояния Хеллингера ρ

$$\mathbf{E}_{P_j} e^{\frac{1}{2} \sum_{n=0}^{\infty} \rho(P_j(y_{1:n}a|y_{1:n}), Q(y_{1:n}a|y_{1:n}))} \leq w_j^{-1/2}. \quad (1)$$

Так как функция e^x выпукла вниз, то

$$e^{\mathbf{E}_{P_j} \frac{1}{2} \sum_{n=0}^{\infty} \rho(P_j(y_{1:n}a|y_{1:n}), Q(y_{1:n}a|y_{1:n}))} \leq \mathbf{E}_{P_j} e^{\frac{1}{2} \sum_{n=0}^{\infty} \rho(P_j(y_{1:n}a|y_{1:n}), Q(y_{1:n}a|y_{1:n}))} \leq w_j^{-1/2}.$$

После логарифмирования и умножения на 2 получается такой же результат, как у Соломонова, только для расстояния ρ , а не ρ' . С другой стороны, мы видим, что мера P_j множества $\{y \mid \sum_{n=0}^{\infty} \rho(P_j(y_{1:n}a|y_{1:n}), Q(y_{1:n}a|y_{1:n})) > s\}$ (которую мы обозначили через $\mathcal{F}_{P_j, Q}(s)$) меньше $w_j^{-1/2} e^{-s/2}$.

Для доказательства (1) Вовк рассматривает вспомогательную меру $R(x)$ на дереве, определяемую из соотношения для условных вероятностей $R(x_{1:n}a|x_{1:n}) = \frac{\sqrt{P_j(x_{1:n}a|x_{1:n})Q(x_{1:n}a|x_{1:n})}}{\sum_b \sqrt{P_j(x_{1:n}b|x_{1:n})Q(x_{1:n}b|x_{1:n})}}$. Пользуясь формулой $\rho(p, q) = 2 - 2 \sum_i \sqrt{p(i)q(i)}$ для расстояния Хеллингера, преобразуем знаменатель:

$$\sum_b \sqrt{P_j(x_{1:n}b|x_{1:n})Q(x_{1:n}b|x_{1:n})} = 1 - \frac{1}{2} \rho(P_j(x_{1:n}b|x_{1:n}), Q(x_{1:n}b|x_{1:n})).$$

Тогда

$$\begin{aligned} R(x_{1:n}) &= \prod_{m=0}^{n-1} \frac{\sqrt{P_j(x_{1:m+1}|x_{1:m})Q(x_{1:m+1}|x_{1:m})}}{\sum_b \sqrt{P_j(x_{1:m}b|x_{1:m})Q(x_{1:m}b|x_{1:m})}} = \\ &= \frac{\sqrt{P_j(x_{1:n})Q(x_{1:n})}}{\prod_{m=0}^{n-1} (1 - \frac{1}{2} \rho(P_j(x_{1:m}b|x_{1:m}), Q(x_{1:m}b|x_{1:m})))}. \end{aligned}$$

Поскольку $e^{-z} \geq 1 - z$, то имеем оценку

$$\begin{aligned} R(x_{1:n}) &\geq \frac{\sqrt{P_j(x_{1:n})Q(x_{1:n})}}{\prod_{m=0}^{n-1} e^{-\frac{1}{2} \rho(P_j(x_{1:m}b|x_{1:m}), Q(x_{1:m}b|x_{1:m}))}} = \\ &= \sqrt{P_j(x_{1:n})Q(x_{1:n})} e^{\frac{1}{2} \sum_{m=0}^{n-1} \rho(P_j(x_{1:m}b|x_{1:m}), Q(x_{1:m}b|x_{1:m}))} = \\ &= \sqrt{\frac{Q(x_{1:n})}{P_j(x_{1:n})}} P_j(x_{1:n}) e^{\frac{1}{2} \sum_{m=0}^{n-1} \rho(P_j(x_{1:m}b|x_{1:m}), Q(x_{1:m}b|x_{1:m}))} \geq \\ &\geq \sqrt{w_j} P_j(x_{1:n}) e^{\frac{1}{2} \sum_{m=0}^{n-1} \rho(P_j(x_{1:m}b|x_{1:m}), Q(x_{1:m}b|x_{1:m}))}. \end{aligned}$$

Таким образом, $w_j^{-1/2} R(x_{1:n}) \geq P_j(x_{1:n}) e^{\frac{1}{2} \sum_{m=0}^{n-1} \rho(P_j(x_{1:m}b|x_{1:m}), Q(x_{1:m}b|x_{1:m}))}$, и после суммирования по всем x длины n получаем (1).

Характеристикой предсказателя назовём точную верхнюю грань тех α , для которых по любым экспертам P_j предсказатель строит такого эксперта Q , что $\forall j \mathcal{F}_{P_j, Q}(s) = O(e^{-\alpha s})$. Предсказатель тем эффективнее, чем больше его характеристика. Для ситуации, когда экспертов двое, ниже получены следующие результаты. Теорема 5 показывает, что характеристика «взвешенного среднего» равна $\frac{1}{2}$, а характеристика «максимального правдоподобия» не превышает $\frac{1}{2}$. В то же время, в теореме 6 построен цельный предсказатель с характеристикой 1. Из теоремы 4 следует, что эту характеристику невозможно увеличить даже в классе всех предсказателей.

Таким образом, предсказатели «взвешенное среднее» и «максимальное правдоподобие» далеко не оптимальны в классе всех предсказателей и в классе всех цельных предсказателей, соответственно. *Характеристикой множества κ* , состоящего из предсказателей, назовём точную верхнюю грань таких α , что для некоторого β и для всех $s > 0$ существует предсказатель из κ , который по любым экспертам P_j строит эксперта Q со свойством $\forall j \mathcal{F}_{P_j, Q}(s) < e^{\beta - \alpha s}$. Характеристика предсказателя очевидно совпадает с характеристикой одноэлементного множества, содержащего этот предсказатель. По теоремам 6 и 3 характеристика класса всех цельных предсказателей равна 1. А вот характеристика класса всех предсказателей оказывается равной 2, в соответствии с теоремами 1 и 2.

4. Точные оценки нескольких характеристик

Лемма 1. Пусть на конечном непустом множестве A дан конечный непустой набор распределений вероятностей $\{p_j\}$. Существует единственное распределение вероятностей q на A , для которого достигает минимума сумма расстояний Хеллингера от p_j до q .

Доказательство. Так как пространство распределений вероятностей на A компактно в обычной метрике, а расстояние Хеллингера непрерывно в обычной метрике, то минимум $\sum_j \rho(p_j, q)$ достигается для какого-то q . Покажем, что если $q(a) = 0$, то $\forall j p_j(a) = 0$. Рассмотрим то b , для которого $q(b) > 0$, и достаточно малое положительное число ε . Определим распределение вероятностей r на A : $r(a) = \varepsilon$, $r(b) = q(b) - \varepsilon$, на остальных элементах A распределение r совпадает с q . Разность между $\sum_j \rho(p_j, r)$ и $\sum_j \rho(p_j, q)$ равна $2(\sqrt{q(b)} - \sqrt{q(b) - \varepsilon}) \sum_j \sqrt{p_j(b)} - 2\sqrt{\varepsilon} \sum_j \sqrt{p_j(a)}$. Поскольку $\sqrt{q(b)} - \sqrt{q(b) - \varepsilon} = O(\varepsilon)$, то в случае $\exists j p_j(a) > 0$ было бы $\sum_j \rho(p_j, r) < \sum_j \rho(p_j, q)$.

Если для некоторого b окажется $q(b) = 1$, то $\forall a \neq b q(a) = 0$, $\forall j \forall a \neq b p_j(a) = 0$ и $\forall j p_j(b) = 1$. Тогда $\sum_j \rho(p_j, q) = 0$ и $\forall r \neq q \sum_j \rho(p_j, r) > 0$.

Пусть B — множество тех элементов A , на которых q положительно и $|B| \geq 2$. Тогда $\sum_{b \in B} q(b) = 1$ и $\forall j \sum_{b \in B} p_j(b) = 1$. Дифференцируя функцию $\sum_j (2 - 2 \sum_{b \in B} \sqrt{p_j(b)} \sqrt{x_b})$ по каждому x_b , получим вектор частных производных, пропорциональный вектору $(1, \dots, 1)$, когда $\forall b x_b = q(b)$. Производная указанной функции по x_b в точке q равна $(-\sum_j \sqrt{p_j(b)}/\sqrt{q(b)})$. То есть $q(b)$ пропорционально $(\sum_j \sqrt{p_j(b)})^2$ и получается $q(b) = \frac{(\sum_j \sqrt{p_j(b)})^2}{\sum_a (\sum_j \sqrt{p_j(a)})^2}$. Последняя формула очевидно выполнена во всех разобранных случаях. \square

Простые вычисления показывают, что

$$\min_q \sum_j \rho(p_j, q) = 2J - 2\sqrt{\sum_a \left(\sum_j \sqrt{p_j(a)} \right)^2},$$

где J — количество распределений p_j . Отметим, что для случая $J = 2$ расстояния $\rho(p_1, q)$ и $\rho(p_2, q)$ совпадают и равны $2 - \sqrt{4 - \rho(p_1, p_2)}$.

Теорема 1. Для каждого $s > 0$ существует предсказатель, который по экспертам P_1, P_2 строит эксперта Q и $\forall j \mathcal{F}_{P_j, Q}(s) < e^2 e^{-2s}$.

Доказательство. Зафиксируем некоторое $s > 0$. Пусть x — последовательность событий. Проходя вдоль x , предсказатель до некоторого шага m выдаёт распределение $Q(x_{1:m}a|x_{1:m})$ на A , минимизирующее сумму расстояний Хеллингера до распределений $P_1(x_{1:m}a|x_{1:m})$ и $P_2(x_{1:m}a|x_{1:m})$. При этом он запоминает величину $c(x_{1:m}) = \sum_{i=0}^m \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i}))$. Ставить в последней формуле P_1 или P_2 — не важно, результат от этого не меняется. Для удобства положим $c(x_{1:-1}) = 0$. Если после шага n первый раз стало $c(x_{1:n}) > s - (2 - \sqrt{2})$ (назовём $x_{1:n+1}$ *точкой переключения*), то предсказатель выбирает j , для которого $P_j(x_{1:n+1})$ больше, и дальше всюду при $m > n$ выдаёт $Q(x_{1:m}a|x_{1:m}) = P_j(x_{1:m}a|x_{1:m})$.

Найдём оценку для $\mathcal{F}_{P_1, Q}(s)$ (в силу симметрии, та же оценка будет верна и для P_2). Заметим, что если $\sum_{i=0}^m \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) > s$, то $m > n$, так как $\sum_{i=0}^{n-1} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) \leq s - (2 - \sqrt{2})$ и $\rho(P_1(x_{1:n}a|x_{1:n}), Q(x_{1:n}a|x_{1:n})) = 2 - \sqrt{4 - \rho(P_1(x_{1:n}a|x_{1:n}), P_2(x_{1:n}a|x_{1:n}))} \leq 2 - \sqrt{2}$. Причём предсказатель в точке $x_{1:n+1}$ переключился на P_2 , а не на P_1 , иначе при $m > n$ имеем $\rho(P_1(x_{1:m}a|x_{1:m}), Q(x_{1:m}a|x_{1:m})) = 0$ и $\sum_{i=0}^m \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) = \sum_{i=0}^n \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) \leq s$. Таким образом, мера P_1 множества $\{x \mid \sum_{i=0}^{\infty} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) > s\}$ не превышает меры P_1 множества M тех точек переключения $x_{1:n+1}$, где $P_2(x_{1:n+1}) \geq P_1(x_{1:n+1})$.

Для оценки $P_1(M)$ используем рассуждения, похожие на рассуждения из доказательства формулы (1). Рассмотрим вспомогательную меру $R(x)$, задаваемую соотношением для условных вероятностей

$$R(x_{1:m}a|x_{1:m}) = \frac{\sqrt{P_1(x_{1:m}a|x_{1:m})P_2(x_{1:m}a|x_{1:m})}}{\sum_b \sqrt{P_1(x_{1:m}b|x_{1:m})P_2(x_{1:m}b|x_{1:m})}}.$$

Заметим, что для произвольных распределений вероятностей p_1, p_2 на A верно неравенство $1 - \frac{1}{2}\rho(p_1, p_2) \leq e^{-2\rho(p_1, q)}$, где q — распределение, минимизирующее сумму расстояний Хеллингера до p_1 и p_2 . (Действительно, если обозначить $\rho(p_1, p_2)$ через ρ , то $\rho(p_1, q) = 2 - \sqrt{4 - \rho}$. Неравенство $\ln(1 - \frac{\rho}{2}) \leq -4 + 2\sqrt{4 - \rho}$ верно при $\rho = 0$, поэтому достаточно его проверить для производных: $\frac{-1}{2-\rho} \leq \frac{-1}{\sqrt{4-\rho}} \Leftrightarrow \sqrt{4-\rho} \geq 2 - \rho \Leftrightarrow \rho(\rho - 3) \leq 0$.) Отсюда по аналогии с доказательством неравенства (1) получаем

оценку для $x_{1:n+1} \in M$

$$\begin{aligned} R(x_{1:n+1}) &= \frac{\sqrt{P_1(x_{1:n+1})P_2(x_{1:n+1})}}{\prod_{i=0}^n (1 - \frac{1}{2}\rho(P_1(x_{1:i}b|x_{1:i}), P_2(x_{1:i}b|x_{1:i})))} \geq \\ &\geq \sqrt{\frac{P_2(x_{1:n+1})}{P_1(x_{1:n+1})}} P_1(x_{1:n+1}) e^{2 \sum_{i=0}^n \rho(P_1(x_{1:i}b|x_{1:i}), Q(x_{1:i}b|x_{1:i}))} \geq P_1(x_{1:n+1}) e^{2s-4+2\sqrt{2}}, \end{aligned}$$

то есть $e^{-2s+4-2\sqrt{2}} R(x_{1:n+1}) \geq P_1(x_{1:n+1})$, и просуммировав по M , получим требуемое. \square

Теорема 2. $\forall \varepsilon > 0$ существуют такие эксперты P_1, P_2 и число s_0 , что $\forall s > s_0$ и для любого эксперта $Q \exists j \mathcal{F}_{P_j, Q}(s) > e^{-(2+\varepsilon)s}$.

Доказательство. Рассмотрим двухэлементное множество исходов $\{0, 1\}$ и достаточно малое $\delta > 0$. Пусть $n > 1/\delta^3$; P_1 -вероятность появления очередного нуля всегда равна $(1 + \delta)/2$, P_1 -вероятность появления очередной единицы всегда равна $(1 - \delta)/2$; P_2 -вероятность появления очередного нуля всегда равна $(1 - \delta)/2$, P_2 -вероятность появления очередной единицы всегда равна $(1 + \delta)/2$. Зафиксируем произвольного эксперта Q .

Обозначим через M множество двоичных слов длины $2n$, в которых ровно n нулей и ровно n единиц. Мощностъ этого множества больше $\frac{2^{2n}}{2\sqrt{n}}$. Мера P_1 и мера P_2 каждого элемента из M равны $\frac{(1-\delta^2)^n}{2^{2n}}$. Поместим в множество M_1 те x из M , для которых

$$\sum_{i=0}^{2n-1} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) \geq \sum_{i=0}^{2n-1} \rho(P_2(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})).$$

Поместим в множество M_2 те x из M , для которых

$$\sum_{i=0}^{2n-1} \rho(P_2(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) \geq \sum_{i=0}^{2n-1} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})).$$

Предположим, что $|M_1| \geq |M|/2$ (предположение $|M_2| \geq |M|/2$ рассматривается симметрично). Для x из M_1 имеем

$$\begin{aligned} &\sum_{i=0}^{2n-1} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) \geq \\ &\geq \frac{1}{2} \left(\sum_{i=0}^{2n-1} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) + \sum_{i=0}^{2n-1} \rho(P_2(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) \right) = \\ &= \sum_{i=0}^{2n-1} \frac{1}{2} (\rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) + \rho(P_2(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i}))). \end{aligned}$$

Как следует из леммы 1,

$$\begin{aligned} &\frac{1}{2} (\rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) + \rho(P_2(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i}))) \geq \\ &\geq 2 - \sqrt{4 - \rho(P_1(x_{1:i}a|x_{1:i}), P_2(x_{1:i}a|x_{1:i}))} > \delta^2(1 - o(1))/4 \end{aligned}$$

(величина $o(1)$ стремится к 0 при $\delta \rightarrow 0$). Таким образом, для x из M_1

$$\sum_{i=0}^{2n-1} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) > n\delta^2(1 - o(1))/2.$$

Мера P_1 множества M_1 равна

$$|M_1| \frac{(1 - \delta^2)^n}{2^{2n}} > e^{-n\delta^2(1+o(1)) - O(\ln n)} > e^{-n\delta^2(1+o(1))}.$$

Положим $s_0 = 1/\delta$. Если $n = \left\lceil \frac{2s}{\delta^2}(1 + o(1)) \right\rceil$ (для подходящей величины $o(1)$), то для $s > s_0$ будет $n > 1/\delta^3$, $n\delta^2(1 - o(1))/2 > s$ и $\mathcal{F}_{P_1, Q}(s) > e^{-(2+\varepsilon)s}$. \square

Теорема 3. $\forall \varepsilon > 0$ существуют такие эксперты P_1, P_2 и число s_0 , что $\forall s > s_0$ и для любого эксперта Q , построенного цельным предсказателем по P_1 и P_2 , $\exists j$ $\mathcal{F}_{P_j, Q}(s) > e^{-(1+\varepsilon)s}$.

Доказательство. Рассуждение проходит аналогично доказательству предыдущей теоремы. Отметим только отличия. Расстояние $\rho(P_j(za|z), Q(za|z))$ теперь равно или 0, или $\delta^2(1 - o(1))$. Следовательно, для x из M_1

$$\sum_{i=0}^{2n-1} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) > n\delta^2(1 - o(1)).$$

Положим значение n на этот раз равным $\left\lceil \frac{s}{\delta^2}(1 + o(1)) \right\rceil$ (для подходящей величины $o(1)$). Тогда будет $n > 1/\delta^3$, $n\delta^2(1 - o(1)) > s$ и $\mathcal{F}_{P_1, Q}(s) > e^{-(1+\varepsilon)s}$. \square

Теорема 4. $\exists \varepsilon_0 > 0 \forall \varepsilon \in (0, \varepsilon_0)$ существуют такие эксперты P_1, P_2 , что $\forall s_1 > 0$ и для любого эксперта Q выполнено хотя бы одно из двух неравенств: $\mathcal{F}_{P_1, Q}(s_1) > e^{-s_1}$, $\mathcal{F}_{P_2, Q}(s_2) > e^{-s_2(1+\varepsilon)}$, где $s_2 = s_1\varepsilon^{-4}$.

Доказательство. Рассмотрим двухэлементное множество исходов $\{0, 1\}$, число $s_1 > 1$ и достаточно малое $\varepsilon > 0$. Пусть P_1 -вероятность появления очередного нуля всегда равна 1, P_1 -вероятность появления очередной единицы всегда равна 0; P_2 -вероятность появления очередного нуля всегда равна $1 - \varepsilon$, P_2 -вероятность появления очередной единицы всегда равна ε . Зафиксируем произвольного эксперта Q .

Положим

$$n = \lceil \varepsilon^{-5}(1 + 4\varepsilon^{1.5})s_1 + 0.3\varepsilon^{-4}s_1 \rceil.$$

Обозначим через M конус над вершиной 0^n . Поскольку его P_1 -мера равна 1, достаточно показать, что из неравенства

$$\sum_{i=0}^n \rho(P_1(0^i a | 0^i), Q(0^i a | 0^i)) \leq s_1$$

следует $\mathcal{F}_{P_2, Q}(s_2) > e^{-(1+\varepsilon)s_2}$.

Если для некоторого i имеет место $Q(0^{i+1}|0^i) < 1 - \frac{10}{3}\varepsilon^4$, то имеет место $\rho(P_1(0^i a|0^i), Q(0^i a|0^i)) > \frac{10}{3}\varepsilon^4$ (мы использовали $\delta \leq 1 \Rightarrow 2 - 2\sqrt{1-\delta} \geq \delta$). Поэтому количество таких i не превышает $0,3\varepsilon^{-4}s_1$. Для остальных же i (которых больше $\varepsilon^{-5}(1 + 4\varepsilon^{1,5})s_1$) получаем

$$\rho(P_2(0^i a|0^i), Q(0^i a|0^i)) > \varepsilon \left(1 - \sqrt{\frac{40}{3}}\varepsilon^{1,5}\right).$$

Действительно, обозначим $Q(0^i 1|0^i)$ через z (по предположению $z \in \left[0, \frac{10}{3}\varepsilon^4\right]$). Тогда

$$\begin{aligned} \rho(P_2(0^i a|0^i), Q(0^i a|0^i)) &= 2 - 2\sqrt{1-\varepsilon}\sqrt{1-z} - 2\sqrt{\varepsilon}\sqrt{z} \geq \\ &\geq 2 - 2\sqrt{1-\varepsilon} - 2\sqrt{\frac{10}{3}}\varepsilon^{2,5} > \varepsilon - \sqrt{\frac{40}{3}}\varepsilon^{2,5}. \end{aligned}$$

Таким образом,

$$\sum_{i=0}^n \rho(P_2(0^i a|0^i), Q(0^i a|0^i)) > s_1\varepsilon^{-4}(1 + 4\varepsilon^{1,5}) \left(1 - \sqrt{\frac{40}{3}}\varepsilon^{1,5}\right) > s_2.$$

Кроме того, P_2 -мера множества M равна

$$\begin{aligned} (1-\varepsilon)^n &> e^{\ln(1-\varepsilon)s_1(\varepsilon^{-5} + 4\varepsilon^{-3,5} + 0,3\varepsilon^{-4} + 1)} > e^{-(\varepsilon + 0,6\varepsilon^2)s_1\varepsilon^{-5}(1 + 0,3\varepsilon + 5\varepsilon^{1,5})} = \\ &= e^{-s_1\varepsilon^{-4}(1 + 0,6\varepsilon)(1 + 0,3\varepsilon + 5\varepsilon^{1,5})} > e^{-s_2(1+\varepsilon)}, \end{aligned}$$

что и доказывает теорему. \square

Теорема 5. $\forall \varepsilon > 0$ существуют такие эксперты P_1, P_2 и число s_0 , что $\forall s > s_0$ если в качестве эксперта Q взять «взвешенное среднее» или «максимальное правдоподобие» от P_1 и P_2 , то $\exists j \mathcal{F}_{P_j, Q}(s) > e^{-(0,5+\varepsilon)s}$.

Доказательство. Рассмотрим двухэлементное множество исходов $\{0, 1\}$, достаточно малое $\delta > 0$ и $h = \lceil \delta^{-3} \rceil$. Пусть для первых h испытаний P_1 -вероятность появления очередного нуля равна 1, а очередной единицы — 0; P_2 -вероятность появления очередного нуля равна $1 - \delta$, а очередной единицы — δ . Для следующих h испытаний, наоборот, P_2 -вероятность появления очередного нуля равна 1, а очередной единицы — 0; P_1 -вероятность появления очередного нуля равна $1 - \delta$, а очередной единицы — δ . Далее условные вероятности экспертов чередуются с периодом $2h$.

Положим $h_0 = \lceil \delta^{-2} \rceil$. Назовём испытание исключительным, если его номер отстоит от какого-нибудь натурального числа кратного $2h$ менее, чем на h_0 . Просто проверяется, что если i -ое испытание неисключительно, то для апостериорных весов выполнено $W_1(0^i) > 1 - \delta^4$ и $W_2(0^i) < \delta^4$ (см. определение предсказателя «взвешенное среднее»). Поэтому $\rho(P_2(0^i a|0^i), Q(0^i a|0^i)) > \delta$. Рассмотрим произвольное $s > \delta^{-3}$ и положим n ближайшим сверху к $\frac{hs}{\delta(h-h_0)}$ кратным $2h$. Обозначим через M конус над вершиной 0^n . Очевидно,

$$\sum_{i=0}^n \rho(P_2(0^i a|0^i), Q(0^i a|0^i)) > s.$$

Кроме того, $n < s\delta^{-1}(1 + O(\delta))$. Следовательно, P_2 -мера множества M равна

$$(1 - \delta)^{n/2} > e^{-\delta(1+O(\delta))s\delta^{-1}(1+O(\delta))/2} > e^{-s(0,5+\varepsilon)},$$

что и доказывает теорему. \square

Теорема 6. *Существует цельный предсказатель, который по экспертам P_1, P_2 строит эксперта Q и $\forall j \forall s \mathcal{F}_{P_j, Q}(s) \leq e^2 e^{-s}$.*

Доказательство. Пусть x — последовательность событий. В каждый момент времени t предсказатель помнит два положительных числа $c_1(x_{1:t}), c_2(x_{1:t})$, таких что $c_1(x_{1:t})c_2(x_{1:t}) = 1$. Эти числа определяются индуктивно. Положим $c_1(x_{1:0}) = c_2(x_{1:0}) = 1$. На каждом шаге предсказатель выдаёт $Q(x_{1:t}a|x_{1:t}) = P_j(x_{1:t}a|x_{1:t})$ для того j , для которого $c_j(x_{1:t})$ меньше (нестрого). Далее, пусть уже выдано какое-то $Q(x_{1:t}a|x_{1:t})$. После этого пересчитываются c_j :

$$c_1(x_{1:t+1}) = c_1(x_{1:t}) \sqrt{\frac{P_2(x_{1:t+1}|x_{1:t})}{P_1(x_{1:t+1}|x_{1:t})}} e^{\pm \frac{1}{2} \rho(P_1(x_{1:t}a|x_{1:t}), P_2(x_{1:t}a|x_{1:t}))},$$

$$c_2(x_{1:t+1}) = c_2(x_{1:t}) \sqrt{\frac{P_1(x_{1:t+1}|x_{1:t})}{P_2(x_{1:t+1}|x_{1:t})}} e^{\pm \frac{1}{2} \rho(P_1(x_{1:t}a|x_{1:t}), P_2(x_{1:t}a|x_{1:t}))},$$

причём знак «+» ставится в показателе экспоненты для того j , для которого $c_j(x_{1:t})$ меньше; для другого j в показателе экспоненты ставится, соответственно, знак «−».

Оценим теперь меру P_1 открытого множества

$$M = \left\{ x \mid \sum_{i=0}^{\infty} \rho(P_1(x_{1:i}a|x_{1:i}), Q(x_{1:i}a|x_{1:i})) > s \right\}$$

(в силу симметрии рассуждения для P_2 будут аналогичны). Теперь будем рассматривать только множество вершин дерева высоты не более n . Обозначим его K_n . Пусть множеству L_n принадлежат те вершины из K_n , конуса над которыми вложены в M . Множество минимальных вершин из L_n обозначим через M_n (если для каждого n оценка меры будет получена для конечного объединения конусов с основаниями из M_n , то она будет получена и для M). Докажем индукцией по убыванию t (от листьев к корню), что для любого $x_{1:t} \in K_n$ выполняется неравенство

$$e^2 c_1(x_{1:t}) \geq \frac{1}{P_1(x_{1:t})} \sum_{\substack{z \in M_n, \\ z_{1:t} = x_{1:t}}} P_1(z) e^{\sum_{i=t}^{|z|-1} \rho(P_1(z_{1:i}a|z_{1:i}), Q(z_{1:i}a|z_{1:i}))}. \quad (2)$$

Если над вершиной $x_{1:t}$ нет вершин из M_n (и сама она не из M_n), то неравенство обращается в очевидно верное: $e^2 c_1(x_{1:t}) \geq 0$ (сумма в правой части пустая). Если $x_{1:t} \in M_n$, то строго над ней нет вершин из M_n , и неравенство принимает вид $e^2 c_1(x_{1:t}) \geq e^{\rho(P_1(x_{1:t}a|x_{1:t}), Q(x_{1:t}a|x_{1:t}))}$. Последнее неравенство следует из того, что $\forall p_1, p_2 \rho(p_1, p_2) \leq 2$ и $c_1(x_{1:t}) \geq 1$ (действительно, так как $x_{1:t}$ — минимальная вершина из тех, в которых сумма расстояний Хеллингера от Q до P_1 стала больше s , то на шаге t было выбрано P_2 , а не P_1 , и значит, $c_1(x_{1:t}) \geq c_2(x_{1:t})$). Таким образом, база индукции доказана.

Теперь проведём индуктивный переход. Пусть известно, что $x_{1:m} \notin M$, и кроме того, для каждого исхода b

$$e^2 c_1(x_{1:m}b) \geq \frac{1}{P_1(x_{1:m}b)} \sum_{\substack{z \in M_n, \\ z_{1:m+1} = x_{1:m}b}} P_1(z) e^{\sum_{i=m+1}^{|z|} \rho(P_1(z_{1:i}a|z_{1:i}), Q(z_{1:i}a|z_{1:i}))}.$$

Тогда правую часть необходимого неравенства (2) можно преобразовать следующим образом (для краткости будем вместо $x_{1:m}$ писать просто x ; распределения вероятностей $P_j(z_{1:i}a|z_{1:i})$, $Q(z_{1:i}a|z_{1:i})$ будем обозначать $p_j(z_{1:i})$ и $q(z_{1:i})$, соответственно):

$$\begin{aligned} \frac{1}{P_1(x)} \sum_{\substack{z \in M_n, \\ z_{1:m} = x}} P_1(z) e^{\sum_{i=m}^{|z|} \rho(p_1(z_{1:i}), q(z_{1:i}))} &= \\ &= e^{\rho(p_1(x), q(x))} \sum_b \frac{1}{P_1(x)} \sum_{\substack{z \in M_n, \\ z_{1:m+1} = xb}} P_1(z) e^{\sum_{i=m+1}^{|z|} \rho(p_1(z_{1:i}), q(z_{1:i}))} = \\ &= e^{\rho(p_1(x), q(x))} \sum_b \frac{P_1(xb|x)}{P_1(xb)} \sum_{\substack{z \in M_n, \\ z_{1:m+1} = xb}} P_1(z) e^{\sum_{i=m+1}^{|z|} \rho(p_1(z_{1:i}), q(z_{1:i}))} \leq \\ &\leq e^{\rho(p_1(x), q(x))} \sum_b P_1(xb|x) e^2 c_1(xb). \end{aligned}$$

Чтобы теперь доказать требуемое неравенство

$$e^{\rho(p_1(x), q(x))} \sum_b P_1(xb|x) c_1(xb) \leq c_1(x), \quad (3)$$

нужно разобрать два случая. Если в вершине x было выбрано $q = p_1$, то $e^{\rho(p_1(x), q(x))} = 1$ и $\forall b$ $c_1(xb) = c_1(x) \sqrt{\frac{P_2(xb|x)}{P_1(xb|x)}} e^{\frac{1}{2} \rho(p_1(x), p_2(x))}$, и (3) обращается в

$$\sum_b \sqrt{P_1(xb|x) P_2(xb|x)} \leq e^{-\frac{1}{2} \rho(p_1(x), p_2(x))}. \quad (4)$$

Доказательство неравенства (4) уже приводилось при доказательстве формулы (1).

Если же в вершине x было выбрано $q = p_2$, то $\forall b$ $c_1(xb) = c_1(x) \sqrt{\frac{P_2(xb|x)}{P_1(xb|x)}} e^{-\frac{1}{2} \rho(p_1(x), p_2(x))}$, и неравенство (3) обращается в

$$e^{\rho(p_1(x), p_2(x))} \sum_b \sqrt{P_1(xb|x) P_2(xb|x)} c_1(x) e^{-\frac{1}{2} \rho(p_1(x), p_2(x))} \leq c_1(x),$$

откуда опять получаем (4).

После того, как неравенство (2) доказано, применим его к корню. Имеем: $e^2 \geq \sum_{z \in M_n} P_1(z) e^{\sum_{i=0}^{|z|} \rho(P_1(z_{1:i}a|z_{1:i}), Q(z_{1:i}a|z_{1:i}))} \geq \sum_{z \in M_n} P_1(z) e^s = P_1(M_n) e^s$, откуда получаем требуемую оценку. \square

5. Открытые вопросы

Естественно, хотелось бы для каждого количества экспертов найти характеристики оптимального предсказателя, оптимального цельного предсказателя, класса всех предсказателей и класса всех цельных предсказателей.

Те же проблемы интересно было бы решить для такого обобщения, когда предсказатель может использовать датчик случайных чисел (в исходном определении, которое изучается в настоящей статье, предсказатели являются детерминированными операторами).

Литература

- [1] *В. Г. Вовк*. Об одной критерии случайности // Доклады АН СССР. 1987. Т. 294, № 6. С. 1298–1302.
- [2] *R. J. Solomonoff*. Complexity-Based Induction Systems: Comparisons and Convergence Theorems // IEEE Transactions on Information Theory. 1978. V. 24, № 4. P. 422–432.
- [3] *M. Li and P. M. B. Vitányi*. An Introduction to Kolmogorov Complexity and its Applications, 2nd edition. Berlin: Springer, 1997.