# ПОСТРОЕНИЕ ОПРЕДЕЛИТЕЛЬНЫХ ТАБЛИЦ ПРИ НЕПОЛНОЙ ИНФОРМАЦИИ О ЧАСТОТАХ ВСТРЕЧАЕМОСТИ ОПРЕДЕЛЯЕМЫХ ОБЪЕКТОВ*

А. А. Федотов
*Институт вычислительных технологий СО РАН*
*Новосибирск, Россия*
e-mail: lesha@adm.ict.nsc.ru

Search trees are intended for the identification of objects in biology, mineralogy etc. The traditional quality criterion of a search tree is the average time of the object identification. It, in its turn, is determined by the probability distribution over the set of objects which is usually not known with certainty. However, some known data about the objects occurrence rate can be used in order to reduce the time consumption for compiling the search trees. The case is considered when the data on the occurrency rate can be represented as some partial order. In this work a method for constructing a search tree close to optimum for thus specified class under the minimax approach is presented. An example of making use of the described algorithm is given.

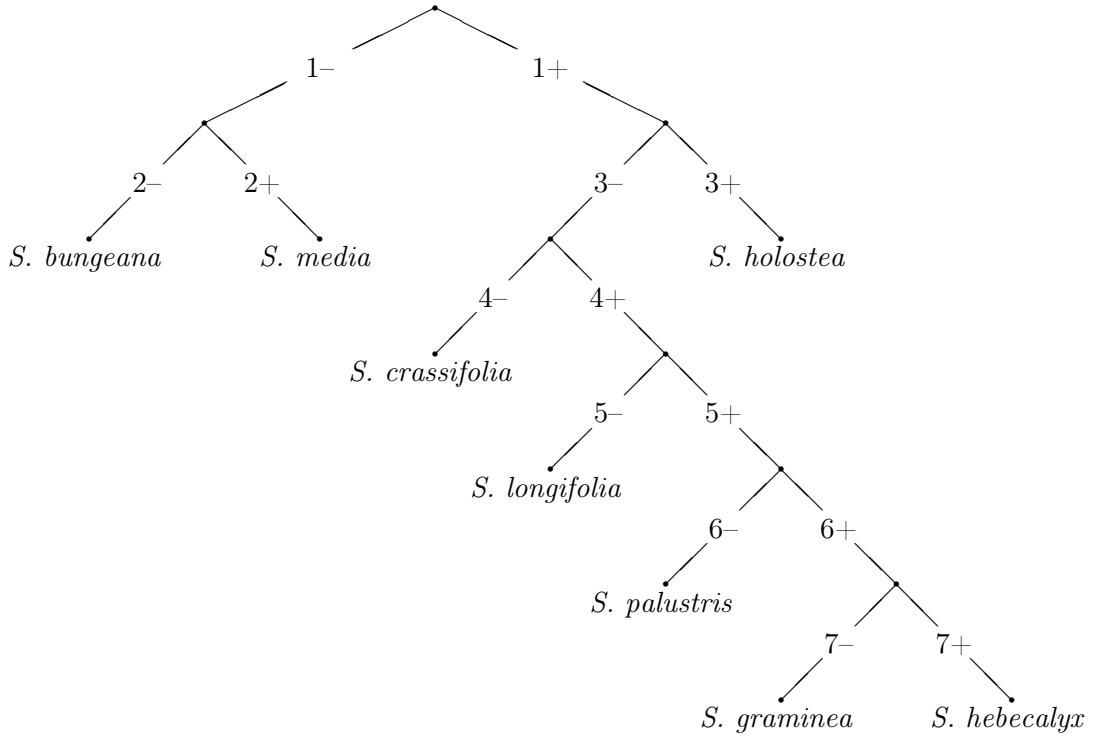
## Введение

В биологии определительные таблицы используются для определения таксономической принадлежности видов. Эта работа отнимает много времени и требует высокой квалификации. Например, при определении насекомых проверка таких типичных признаков, как число и расположение жилок на крыльях, предполагает использование микроскопов и бинокуляров. Поэтому весьма важной представляется задача построения оптимальных ключей, т. е. таких ключей, использование которых занимает в среднем минимальное время. Один из подходов к решению этой задачи будет представлен в настоящей статье.

Определительную таблицу можно представить как двоичное дерево, листьям которого сопоставлены названия, а развилкам — признаки объектов. На рис. 1 изображено определительное дерево для рода *Stellaria* — Звездчатка семейства гвоздичных, полученное ограничением ключа из [9] . Этот род необычайно богат видами и подвидами, поэтому мы ограничились лишь теми, которые встречаются на территории Новосибирской области.

Определение таксономической принадлежности вида с помощью ключа можно представить как движение от корня дерева к одному из листьев. На каждом шаге проверяется

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, грант №98–01–00772.




| | |
|---|---|
| 1– | Листья, по крайней мере нижние, отчетливо черешковые, широкояйцевидные, продолговатые или яйцевидно-серповидные. |
| 1+ | Листья сидячие, эллиптические, ланцетные, линейные или линейно-ланцетные. |
| 2– | Стебли опушены равномерно многоклеточными простыми и железистыми волосками. Чашелистики продолговатые, опушенные по всей поверхности, лепестки в 1.5 – 2 раза длиннее чашелистиков. |
| 2+ | Стебли с одной линией курчавых волосков на междоузлиях. Чашелистики яйцевидно-ланцетные, голые или опушенные в верхней или нижней части. |
| 3– | Листья эллиптические, ланцетные или линейные, 0.5 – 3 см длины. |
| 3+ | Листья ланцетные, длинно заостренные, 4 – 9 см длины, по краю и средней жилке жесткореснитчатые. |
| 4– | Листья эллиптические или широколанцетнолинейные, 3 – 30 мм длины, 1 – 20 мм ширины, длина обычно превышает ширину в 1.5 – 3 раза. |
| 4+ | Листья линейные или узколанцетные, 3 – 40 мм длины, 1 – 8 мм ширины, длина обычно превышает ширину более чем в 4 раза. |
| 5– | Стебли по всей поверхности и листья по краю с обильными мелкими шипиками. |
| 5+ | Стебли гладкие голые или опушенные, иногда только в узлах либо с редкими мелкими шипиками, но тогда листья с ресничками у основания. |
| 6– | Чашелистики целиком голые, листья и стебли обычно голые, реже с мелкими шипиками, сизые или зеленые. |
| 6+ | Чашелистики опушенные или голые, но по краю с ресничками, листья в основании с ресничками, реже голые. |
| 7– | Соцветие раскидистое, многоцветковое (более 10 цветков), прицветники обычно реснитчатые, изредка голые. |
| 7+ | Соцветие б. м. сжатое, малоцветковое (менее 10 цветков), прицветники обычно голые, изредка у верхушки реснитчатые. |

Рис. 1. Определительная таблица для рода Stellaria (звездчатка).

признак в текущей развилке, и дальнейшее направление движения выбирается в зависимости от результата проверки. Например, в примере на рис. 1 необходимо перейти к проверке второго или третьего признака в зависимости от формы листьев экземпляра, видовая принадлежность которого устанавливается.

Известно, что представители различных видов встречаются неравномерно. Тогда естественно считать, что в оптимальном ключе для определения часто встречающихся видов должно использоваться меньше признаков, чем для определения редких видов. Точнее, в теории информации известно, что оптимальным является соотношение, когда среднее время определения каждого вида пропорционально логарифму частоты его встречаемости.

К сожалению, задача построения оптимальной определительной таблицы осложняется еще и тем, что вероятностное распределение, от которого сильно зависит среднее время определения вида и, следовательно, качество определительного дерева, как правило, не известно точно. Более того, узнать его практически невозможно, так как оно сильно варьируется в зависимости от сезона и биотопа. В лучшем случае удается ограничить множество возможных вероятностных распределений.

Например, у биологов имеются некоторые сведения о частотах встречаемости видов. Так, представители некоторых редких видов встречаются $2-3$ раза за сезон, тогда как другие — много раз за день. В таком случае можно достоверно утверждать, что первые встречаются реже. Эти сведения можно задать в виде неравенств между частотами встречаемости видов. Таким образом, мы получим частичный порядок на множестве частот встречаемости. В данной работе рассмотрена задача построения эффективной определительной таблицы в случае, когда сведения о частотах встречаемости видов представлены в виде такого частичного порядка. Это достаточно естественная постановка задачи, так как обычно информация о частотах встречаемости, публикуемая в биологических источниках, может быть представлена в таком виде.

При этом возникает задача оценки качества не на одном вероятностном распределении, а на целом классе. Как показано в литературе по теории информации, среднее время определения вида не является в данном случае удачной мерой качества [1]. Обычно за разумную меру качества на классе вероятностных распределений берут максимум так называемой избыточности дерева по всем возможным вероятностным распределениям. Оказывается, деревья, близкие к оптимальным в этом смысле, можно эффективно строить [4], используя методы теории так называемого универсального кодирования [1]. В настоящей работе мы опишем алгоритм, решающий задачу построения определительного дерева, близкого к оптимальному в вышеназванном смысле, т. е. алгоритм, по заданной информации о частичном порядке на множестве частот видов строящий определительное дерево, качество которого отличается от минимально возможного не более чем на единицу.

Автор выражает благодарность научному руководителю Б. Я. Рябко за постановку данной задачи и многочисленные советы и замечания, которые существенно улучшили качество статьи.

## 1. Математическая постановка задачи

Выразим количественно оценку эффективности определительной таблицы. Предположим, что время проверки всех признаков одинаково. Будем отождествлять множество из $m$ видов с множеством натуральных чисел $\{1, \ldots, m\}$. Тогда, если на этом множестве зада-

но вероятностное распределение $P = \{p_i\}_{i=1}^m$, то среднее время определения вида равно $C(P, L) = \sum_{i=1}^m p_i L_i$, где $L_i$ — длина пути в двоичном дереве от корня до соответствующего листа, равная числу проверяемых признаков. Легко показать [1], что данная величина не меньше энтропии вероятностного распределения $H(P) = -\sum_{i=1}^m p_i \log p_i$. Под log здесь и далее мы понимаем двоичный логарифм, а под $\exp_2$ — обратную ему показательную функцию. Поэтому если необходимо определять эффективность дерева $L$ на нескольких вероятностных распределениях, в качестве оценки берут избыточность $R(P, L) = C(P, L) - H(P)$.

Для примера рассмотрим определительное дерево на рис. 1. Если бы представители всех видов, указанных на этом рисунке, встречались с одинаковой частотой, то среднее время определения по этому дереву было бы равно $\frac{1}{8}(2+2+2+3+4+5+6+6) = 3.75$. Это на 0.75 больше энтропии равномерного распределения, равной $H(P) = \log 8 = 3$. Таким образом, избыточность этого дерева при данном вероятностном распределении равна 0.75. С другой стороны, в табл. 1 указано вероятностное распределение, на котором избыточность рассматриваемого дерева равна нулю.

Таблица 1

Распределение вероятностей, минимизирующее избыточность дерева с рис. 1

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *S. bungeana* | $2^{-2}$ | *S. media* | $2^{-2}$ | *S. holostea* | $2^{-2}$ | *S. crassifolia* | $2^{-3}$ |
| *S. longifolia* | $2^{-4}$ | *S. palustris* | $2^{-5}$ | *S. graminea* | $2^{-6}$ | *S. hebecalyx* | $2^{-6}$ |

Обозначим через $\mathcal{P} \in \mathbb{R}^m$ множество возможных вероятностных распределений, заданное частичным порядком на множестве частот видов. Определим избыточность дерева на классе $\mathcal{P}$ как его избыточность в наихудшем случае $R(\mathcal{P}, L) = \sup_{P \in \mathcal{P}} R(P, L)$.

Теперь мы можем точно сформулировать решаемую задачу: по заданному множеству $\mathcal{P}$ построить дерево $L$, минимизирующее величину $R(\mathcal{P}, L)$. В статье приведен алгоритм, позволяющий приближенно (с точностью до единицы) найти минимум этой функции и дерево, на котором он достигается.

## 2. Алгоритм

Опишем алгоритм для решения вышепоставленной задачи. По множеству $\mathcal{P} \subset \mathbb{R}^m$ возможных вероятностных распределений, заданному соотношениями $\sum_{i=1}^m p_i = 1$, $p_i \geq 0$ при $i = 1, \ldots, m$ и частичным порядком $p_i \leq p_j$ при $(i, j) \in I$, надо построить дерево $L$, минимизирующее величину $R(\mathcal{P}, L)$.

Основная идея алгоритма — сведение нашей задачи к известной в теории информации задаче о вычислении пропускной способности канала, что позволит использовать известный алгоритм Блэхута — Аримото [6]. Для этого потребуется следующая теорема о минимаксе.

**Теорема 1.** *Предположим, что $X$ и $Y$ — выпуклые компактные подмножества линейного пространства. Пусть $f(x, y) : X \times Y \to \mathbb{R}$ — непрерывная функция, вогнутая по $x$ и выпуклая по $y$ (выпуклость и вогнутость функций понимается как соответствующее свойство надграфиков этих функций). Тогда*

$$\inf_{y \in Y} \sup_{x \in X} f(x, y) = \sup_{x \in X} \inf_{y \in Y} f(x, y).$$

Доказательство этого факта в данной формулировке можно найти в [11].

Алгоритм можно условно подразделить на три шага. Первый шаг — построение по исходным данным некоторого канала связи. Поясним, как это сделать. Заметим, что $\mathcal{P}$ как непустое ограниченное множество, заданное системой линейных неравенств, представляет собой выпуклый многогранник. Обозначим множество вершин этого многогранника через $V(\mathcal{P})$. Заметим, что каждая точка $P_i \in \mathcal{P}$ задает вероятностное распределение $\{Q_{ij}\}_{j=1}^m$ на множестве $\{1,\dots,m\}$. Рассмотрим канал $Q$, выходами которого являются элементы $\{1,\dots,m\}$, а входами — точки из $V(\mathcal{P})$. Матрицей канала будем называть набор чисел $\{Q_{ij}\}_{i=1,\,j=1}^{n,\,m}$, где $n = \operatorname{Card} V(\mathcal{P})$.

Вероятностному распределению $q = \{q_i\}_{i=1}^n$ на входе канала соответствует распределение $p = Qq = \{p_j\}_{j=1}^m$ на выходе. Рассмотрим взаимную информацию между этими распределениями — меру зависимости этих распределений:

$$I(q,p) = \sum_{i=1,\,j=1}^{n,\,m} \omega_{ij} \log \frac{\omega_{ij}}{q_i p_j},$$

где $\omega_{ij}$ — совместное распределение на входе и выходе. В нашем случае $\omega_{ij} = q_i Q_{ij}$, и следовательно,

$$I(q,Q) = \sum_{i=1,\,j=1}^{n,\,m} q_i Q_{ij} \log \frac{Q_{ij}}{p_j}.$$

Пропускной способностью канала $Q$ называют максимальную взаимную информацию между входом и выходом, т. е. $\sup_{q\in\Omega} I(q,Q)$, где $\Omega$ — множество всех возможных вероятностных распределений на входе канала.

Заметим, что функцию $I(q,Q)$ можно по непрерывности определить для вырожденных каналов и вероятностных распределений, так как $\lim\limits_{x\to 0} x\log x = 0$. Отсюда следует, что пропускная способность канала как максимум непрерывной функции на компакте достигается на некотором распределении $q \in \Omega$.

Теперь можно сформулировать следующую теорему из работы [4]. Для связности изложения, мы докажем ее в том частном случае, который потребуется нам для обоснования алгоритма.

**Теорема 2.** *Построим по множеству угловых точек $V(\mathcal{P})$ множества $\mathcal{P}$ матрицу $Q$. Пусть каждая строчка в этой матрице $Q$ — это строка координат соответствующей угловой точки. Рассмотрим одноименный канал, соответствующий матрице $Q$. Тогда пропускная способность $C(Q)$ этого канала равна минимуму величины $R(\mathcal{P},L)$ с точностью до единицы. Более точно,*

$$C(Q) \le \inf_{L\in\mathcal{L}} R(\mathcal{P},L) < C(Q) + 1,$$

*где $\mathcal{L}$ — множество всех возможных двоичных деревьев.*

**Доказательство.** Так как функция $R(P,L) = \sum_{i=1}^m (L_i + \log p_i)p_i$ выпукла по $P$, то она достигает максимума в одной из вершин $V(\mathcal{P})$ многогранника $\mathcal{P}$. Таким образом,

$$R(\mathcal{P},L) = \sup_{P\in V(\mathcal{P})} R(P,L).$$

Отметим, что отсюда очевидно следует, что этот супремум являлся бы непрерывной (и кусочно-линейной) функцией аргумента $L$, не будь $\mathcal{L}$ дискретным множеством. Последнюю проблему легко обойти. Расширим множество $\mathcal{L}$ до подмножества большого куба $\tilde{\mathcal{L}} \subset [0;l]^m$, точки которого удовлетворяют неравенству $\sum_{i=1}^m 2^{-L_i} \le 1$.

Заметим, что $C(Q)$ является непрерывной функцией $Q_{ij}$ на множестве всех возможных информационных каналов. Функция $R(\mathcal{P}, L)$, как максимум значений линейных функций на множестве $V(\mathcal{P})$ также непрерывно изменяется при изменении этого множества. Таким образом, без ограничения общности можно считать, что все $Q_{ij} \geq \epsilon$ для некоторого $\epsilon > 0$. Остальные случаи получаются предельным переходом. Тогда можно положить сторону куба $l$ равной $-\log \epsilon$.

Далее покажем, что $C(Q) = \inf_{L \in \tilde{\mathcal{L}}} R(\mathcal{P}, L)$. Отсюда $C(Q) \leq \inf_{L \in \mathcal{L}} R(\mathcal{P}, L)$ будет выполнено в силу того, что $\mathcal{L} \subset \tilde{\mathcal{L}}$. С другой стороны, если $\tilde{L} \in \tilde{\mathcal{L}}$ — точка, на которой достигается инфинум $R(V(\mathcal{P}), L)$, то рассмотрев точку $\left\lceil \tilde{L} \right\rceil \in L$ с координатами $\left\lceil \tilde{L}_i \right\rceil$, получим $\inf_{\mathcal{L}} R(\mathcal{P}, L) \leq R(\mathcal{P}, \left\lceil \tilde{L} \right\rceil) < C(Q) + 1$.

Теперь расширим множество $V(\mathcal{P})$ так, чтобы можно было говорить о непрерывности $R(P, L)$. Рассмотрим множество вероятностных распределений $\Omega$ на множестве $V(\mathcal{P})$. Легко видеть, что $\sup_{P \in V(\mathcal{P})} R(P, L) = \sup_{q \in \Omega} R(q, L)$, где $R(q, L) = \int_{V(\mathcal{P})} R(P, L)\, dq$. Это следует из того, что функция $R(P, L)$ достигает максимума в некоторой точке $P_0 \in V(\mathcal{P})$. Таким образом, супремум правой части достигается на вероятностном распределении, сосредоточенном в точке $P_0$.

Заметим, что как пересечение двух выпуклых множеств $\tilde{\mathcal{L}}$ выпукло, а множество $\Omega$ можно представлять в виде единичного симплекса в линейном пространстве размерности $\operatorname{Card} V(\mathcal{P})$. Более того, функция $R(q, L)$ линейна по обоим своим аргументам и, следовательно, удовлетворяет условиям теоремы о минимаксе. Таким образом,

$$\inf_{L \in \tilde{\mathcal{L}}} \sup_{q \in \Omega} R(q, L) = \sup_{q \in \Omega} \inf_{L \in \tilde{\mathcal{L}}} R(q, L).$$

Осталось, применив метод множителей Лагранжа, найти

$$\inf_{L \in \tilde{\mathcal{L}}} R(q, L) = \inf_{L \in \tilde{\mathcal{L}}} \sum_{i=1}^{n} \left( \sum_{j=1}^{m} q_i Q_{ij} L_j - q_i H(P_i) \right).$$

Хорошо известно, что этот infinum достигается при $L_j = -\log p_j$, где $p_j = \sum\limits_{i=1}^{n} (q_i Q_{ij})$ — распределение на выходе канала. Следовательно,

$$\inf_{L \in \tilde{\mathcal{L}}} R(\mathcal{P}, L) = \sup_{q \in \Omega} \inf_{L \in \tilde{\mathcal{L}}} R(q, L) = \sup_{q \in \Omega} \sum_{i=1,\, j=1}^{n,\, m} (-q_i Q_{ij} \log p_j + q_i Q_{ij} \log Q_{ij}),$$

где правая часть является одним из многих эквивалентных определений пропускной способности канала $Q$, что и требовалось доказать.

Теперь можно изложить общую схему алгоритма:

1) сначала по заданному частичному порядку строим множество граничных точек $V(\mathcal{P})$ и матрицу $Q$ соответствующего канала связи;

2) с помощью алгоритма Блэхута — Аримото [6] находим пропускную способность канала $Q$ и вероятностное распределение $q^*$ на множестве входов $V(\mathcal{P})$, при котором эта пропускная способность достигается;

3) зная вероятностное распределение $q^*$, находим искомое дерево $L^*$, минимизирующее функцию $R(\mathcal{P}, L)$ с точностью до единицы. Для этого надо заметить, что $L^*$ — оптимальное дерево для расределения $p_0 = Qq_0$ на выходе канала.

Приведем теоремы, обосновывающие и поясняющие три шага алгоритма.

**Теорема 3.** *Многогранник $\mathcal{P}$ в пространстве $\mathbb{R}^n$, заданный уравнением $\sum_{i=1}^m p_i = 1$ и системой неравенств $p_i \geq 0$ при всех $i = 1 \dots m$, $p_i \leq p_j$ при $(i,j) \in I$, может быть задан как выпуклая линейная оболочка следующих точек:*

$$V(\mathcal{P}) = \{(a_1/s, \dots, a_m/s) | \{a_i\}_{i=1}^m \in \{0,1\}^m, \;\; a_i \leq a_j \;\; \textit{при} \;\; (i,j) \in I, \; s = a_1 + \dots + a_m\}.$$

**Доказательство.** Очевидно, что все указанные точки лежат в данном многограннике. Докажем индукцией по размерности пространства, что их выпуклая оболочка совпадает с многогранником $\mathcal{P}$.

База индукции очевидна: при $m = 1$ многогранник состоит из единственной точки $p_1 = 1$, которая и составляет множество $V(\mathcal{P})$.

Докажем шаг индукции. Предположим, что нам задана точка $p = (p_1, \dots, p_n)$. Если одна из координат этой точки равна нулю, то задача редуцируется к задаче меньшей размерности. При этом из множества $V(\mathcal{P})$ используются только те точки, у которых соответствующая координата равна нулю. Если же все координаты точки $p$ ненулевые, то рассмотрим наименьшую из них $p_x$. Так как в множестве $V(\mathcal{P})$ всегда есть точка $q = (1/m, \dots, 1/m)$, то представим $p$ в виде выпуклой линейной комбинации этой точки $q$ и некоторой точки $r$, как $mp_x q + (1 - mp_x)r$. При этом у точки $r$ одна из координат уже будет равна нулю, и следовательно, $r$ будет представляться в виде выпуклой линейной комбинации точек из $V(\mathcal{P})$ по предположению индукции, что и требовалось доказать.

Используя эту теорему, легко выписать все граничные точки многогранника заданного частичным порядком. В любом случае этот алгоритм требует порядка $2^n$ операций. Однако этот алгоритм можно несколько улучшить, используя индуктивное построение набора точек в пространстве $\{0,1\}^m$ из набора точек в пространстве $\{0,1\}^{m-1}$. Таким образом мы сможем построить канал $Q$. Приведем утверждения из статьи [6], позволяющие вычислить его пропускную способность.

**Теорема 4.** *Рассмотрим $n \times m$ канал $Q$. Для любого $m \times n$ канала $P$ положим*

$$J(q, Q, P) = \sum_{i=1,\, j=1}^{n,\, m} q_i Q_{ij} \log \frac{P_{ji}}{q_i}.$$

*Тогда выполнены следующие утверждения:*

*1) пропускная способность канала $C$ равна $\max_q \max_P J(q, Q, P)$;*

*2) при фиксированном $q$ максимум $J(q, Q, P)$ достигается при*

$$P_{ji} = \frac{q_i Q_{ij}}{\sum_{i'} q_{i'} Q_{i'j}}; \tag{1}$$

*3) при фиксированном $P$ максимум $J(q, Q, P)$ достигается при*

$$q_i = \frac{\exp_2 \left( \sum_j Q_{ij} \log P_{ji} \right)}{\sum_{i'} \exp_2 \left( \sum_j Q_{i'j} \log P_{ji'} \right)}, \tag{2}$$

*где* $\exp_2(x) = 2^x$.

**Доказательство.** Для доказательства первого пункта теоремы достаточно показать, что взаимная информация $I(q, Q) = \max_P J(q, Q, P)$. Положим

$$p_j = \sum_{i=1}^{n} q_i Q_{ij}, \qquad P^*_{ji} = \frac{q_i Q_{ij}}{p_j}.$$

Легко видеть, что выбранные таким образом значения $P^*_{ji}$ задают $m \times n$ канал связи. Тогда $I(q,Q) = J(q,Q,P^*)$. Покажем, что $J(q,Q,P^*) - J(q,Q,P) \geq 0$, причем равенство достигается только при $P = P^*$:

$$J(q,Q,P^*) - J(q,Q,P) = \sum_{i=1,\,j=1}^{n,m} p_j P_{ji} \log \frac{P^*_{ji}}{P_{ji}} \geq \log e \left( \sum_{i=1,\,j=1}^{n,m} p_j P^*_{ji} - \sum_{i=1,\,j=1}^{n,m} p_j P_{ji} \right) = 0.$$

При доказательстве неравенства мы воспользовались неравенством $\ln x \geq 1 - 1/x$, обращающимся в равенство при $x = 1$. Заметим, что мы тоже доказали формулу (1).

Для доказательства (2) воспользуемся методом множителей Лагранжа. В качестве функции Лагранжа возьмем $\Lambda(q_i,\lambda) = \sum_{i=1,\,j=1}^{n,m} q_i Q_{ij} \log(P_{ji}/q_i) + \lambda \left(\sum_{i=1}^{n} q_i - 1\right)$. Тогда

$$\begin{aligned} \frac{\partial \Lambda(q_i,\lambda)}{\partial q_i} &= -\log q_i - \log e + \sum_{j=1}^{m} Q_{ij} \log P_{ji} + \lambda = 0, \\ q_i &= \mu \exp_2 \left( \sum_{j=1}^{m} Q_{ij} \log P_{ji} \right). \end{aligned}$$

Множитель $\mu$ выбирается из условия $\sum_{i=1}^{n} q_i = 1$. Отсюда следует доказываемое.

Подставив (1) в (2), легко получить следующее следствие.

**Следствие 1.** *Если пропускная способность канала достигается на вероятностном распределении $q$, то $q_i = q_i c_i / c$, где*

$$c_i = \exp_2 \left( \sum_{j=1}^{m} Q_{ij} \log \frac{Q_{ij}}{\sum_{i'=1}^{n} q_{i'} Q_{i'j}} \right), \qquad c = \sum_{i=1}^{n} q_i c_i.$$

Теперь можно объяснить, как работает алгоритм Блэхута — Аримото (рис. 2). Сначала выбираются некоторые произвольные $q^0$ и $P^0$, а затем максимум $J(q,Q,P)$ ищется последовательными уточнениями этих начальных приближений по формулам из теоремы 4. Прежде всего $J(q,Q,P)$ максимизируется по $P$ при постоянном $q$, а затем по $q$ при постоянном $P$. В работе [6] доказано, что при этом величина $J(q,Q,P)$ сходится к пропускной способности канала, а распределение $q$ — к искомому распределению $q^*$ на входе. Так как значение $P$, на котором достигается максимум $J(q,Q,P)$, нам не нужно, предлагается сразу вычислять новое значение $q$ по формуле из следствия 1.

Для формулировки критерия остановки алгоритма Блэхута — Аримото окажется полезным следующее следствие.

**Следствие 2.** *Пропускная способность канала достигается на вероятностном распределении $q$ тогда и только тогда, когда существует такое число $c$, что $c_i = c$ при $q_i \neq 0$ и $c_i \leq c$ при $q_i = 0$, где*

$$c_i = \exp_2 \left( \sum_{j=1}^{m} Q_{ij} \log \frac{Q_{ij}}{\sum_{i'=1}^{n} q_{i'} Q_{i'j}} \right).$$

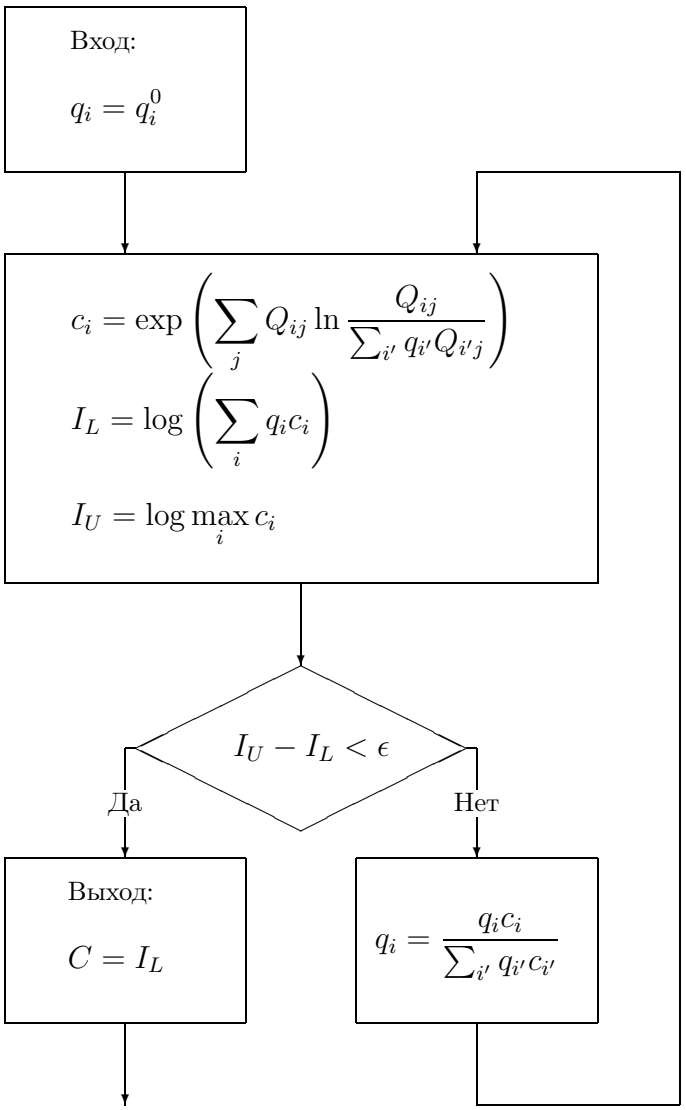



Рис. 2. Алгоритм вычисления пропускной способности канала.

**Доказательство.** В книге [7, раздел 4.4] приведено доказательство того, что функция $I(q, Q)$ выпукла вверх. Там же указан следующий критерий максимума функции, выпуклой вверх на единичном симплексе.

**Теорема 5 (условия Куна — Такера).** *Пусть $f(\alpha)$ является выпуклой вверх функцией в области $\Omega \subset \mathbb{R}^n$, где $\Omega$ — единичный симплекс. Предположим, что частные производные $\partial f(\alpha)/\partial\alpha_k$ определены и непрерывны в области $\Omega$ с тем возможным исключением, что $\lim_{\alpha_k \to 0} \partial f(\alpha)/\partial\alpha_k = +\infty$ для некоторых $k$. Тогда следующие условия являются необходимыми и достаточными условиями того, что в $\alpha$ достигается максимум $f$:*

$$\frac{\partial f(\alpha)}{\partial \alpha_k} = C \text{ при } \alpha_k > 0, \quad \frac{\partial f(\alpha)}{\partial \alpha_k} \le C \text{ при } \alpha_k = 0 \tag{3}$$

*при некотором $C$.*

Для доказательства следствия 2 достаточно проэкспоненцировать условия 3 из теоремы 5 в применении к функции $I(q, Q)$.

Теорема 6 позволит найти искомое дерево по найденному распределению $q^*$ на входе канала.

**Теорема 6.** *Дерево, на котором достигается минимум $R(\mathcal{P}, L)$ с точностью до единицы, — это оптимальное дерево для вероятностного распределения $p^*$ на выходе канала, соответствующего найденному распределению $q^*$ на входе.*

**Доказательство.** Воспользуемся следующим следствием теоремы о минимаксе. Если $\sup_{q\in\Omega} \inf_{L\in\tilde{\mathcal{L}}} R(q, L)$ достигается при $q = q^*$, $L = L^*$, то и $\inf_{L\in\tilde{\mathcal{L}}} \sup_{q\in\Omega} R(q, L)$ достигается там же. Такая пара $(q^*, L^*)$ называется седловой точкой функции $R(q, L)$.

Найдем сначала минимум $R(\mathcal{P}, L)$ на множестве $\tilde{\mathcal{L}}$. Из доказательства теоремы 2 видно, что $\sup_{q\in\Omega} \inf_{L\in\tilde{\mathcal{L}}} R(q, L)$ есть ни что иное, как пропускная способность канала $Q$. Таким образом, наша задача — по известному распределению на входе канала $q^*$ найти $L^*$ как $\arg\inf_{L\in\tilde{\mathcal{L}}} R(q^*, L)$. Рассмотрим распределение $p^* = Qq^*$ на выходе канала $Q$. Легко видеть, что $R(p^*, L)$ отличается от $R(q^*, L)$ на константу, не зависящую от $L$. Таким образом, $L_j^* = -\log p_j^*$.

Для завершения доказательства осталось заметить, что дерево $\lceil L^* \rceil$ с длинами ветвей $\lceil L^* \rceil_j = \lceil -\log p_j^* \rceil$ дает минимум $R(\mathcal{P}, L)$ с точностью до единицы.

# 3. Пример работы алгоритма

Рассмотрим построение определительной таблицы для представителей рода *Stellaria* — Звездчатка, встречающихся на территории Новосибирской области. По данным [5] можно составить табл. 2, содержащую информацию о частотах встречаемости представителей различных видов этого таксона.

Данные о частотах встречаемости можно выразить в виде системы неравенств:

$$\begin{cases} p_1 \le p_2, \\ p_2 \le p_j \quad \text{при} \quad j = 3, \ldots, 8. \end{cases} \tag{4}$$

Таблица 2

Частоты встречаемости различных видов рода *Stellaria* — Зведчатка

| Название вида | Частота | |
|---|---|---|
| *S. hebecalyx* | исчезает | $p_1$ |
| *S. holostea* | редкий | $p_2$ |
| *S. bungeana* | нередкий | $p_3$ |
| *S. crassifolia* | нередкий | $p_4$ |
| *S. graminea* | нередкий | $p_5$ |
| *S. longifolia* | нередкий | $p_6$ |
| *S. media* | нередкий | $p_7$ |
| *S. palustris* | нередкий | $p_8$ |

Этой системе будет соответствовать следующий канал связи:

$$Q = \begin{pmatrix} 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 \\ 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

Заметим, что согласно теореме 3 мы получили бы значительно больше граничных точек. В данном случае их количество можно значительно уменьшить. Подробнее этот вопрос исследуется в работе [12].

В данном примере возможно аналитически вычислить распределение вероятностей, на котором достигается пропускная способность канала. Положим $p_i = p = 1/(6 + 2/8^4)$ при $i = 3, \ldots, 8$, а $p_1 = p_2 = p/8^4$. Ограничимся проверкой того, что это вероятностное распределение реализует пропускную способность канала. Для этого проверим условия Куна — Такера (теорема 5) для данного вероятностного распределения. Вычислим значения $\log c_\alpha$ по формуле

$$\log c_\alpha = \sum_{j=1}^{m} Q_{\alpha j} \log \frac{Q_{\alpha j}}{p_j}.$$

В результате получим, что все $c_\alpha$, за исключением $c_2$, равны $1/p$. Но так как вероятность $q_2$ на входе канала равна нулю, то достаточно показать, что $\log c_2 \leq -\log p$. Действительно,

$$\frac{1}{7} \log \frac{1/7}{p_1} + \frac{6}{7} \log \frac{1/7}{p} = \frac{1}{7} \log 8^4 + \log \frac{1}{7} - \log p = \frac{18}{7} - \log 7 - \log p < -\log p,$$

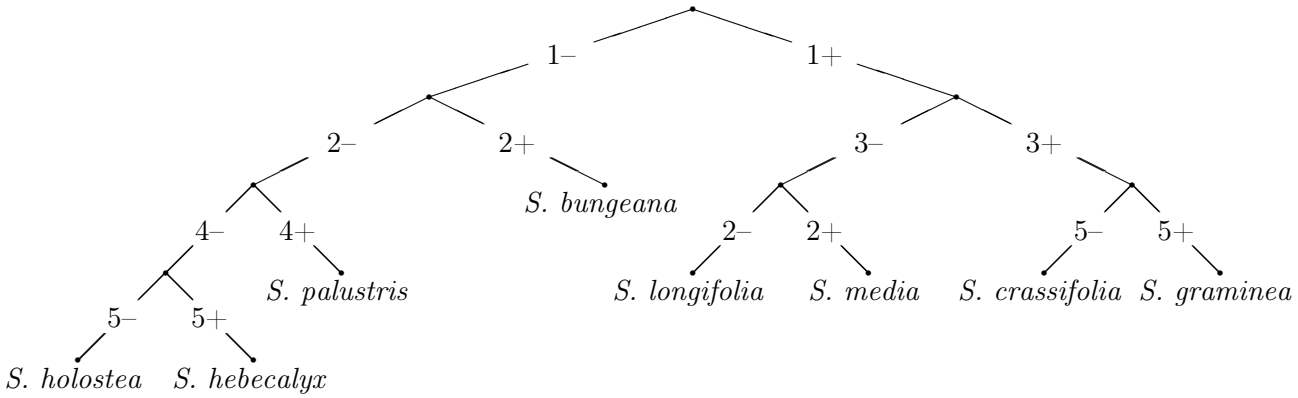
так как $2^{18} < 7^7$.

Найденные вероятности находятся в соответствии с результатами численных расчетов. Теперь легко построить оптимальное дерево, воспользовавшись, например, алгоритмом Хаффмана. В итоге получим дерево, изображенное на рис. 3.

Симплекс-методом легко найти максимум трудоемкости приведенных деревьев на множестве вероятностных распределений, заданном неравенствами 4. Для дерева, изображенного на рис. 1, этот максимум равен 6 и достигается на точечном распределении, сосредоточенном на виде *S. graminea*. Легко видеть, что в данном случае максимум избыточности этого дерева также равен 6.

Для дерева, найденного с использованием нашего алгоритма, максимум трудоемкости достигается на равномерном распределении и равен $3\frac{1}{8}$. Максимум избыточности легко найти, вспомнив, что функция избыточности выпукла вниз на множестве $\mathcal{P}$. Для этого достаточно вычислить ее во всех угловых точках. Максимум, равный 3, достигается на точечном вероятностном распределении, сосредоточенном, например, на том же виде *S. graminea*. Таким образом, по выбранному критерию качества построенное дерево эффективнее на три признака.

В заключение остановимся на возможностях практического применения алгоритма. Наши эксперименты показали, что при числе видов около полутора десятков вычисления




| №п/п | Признак | Таксон | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | a | b | c | d | e | f | g | h |
| 1 | Коробочка существенно длиннее чашечки | − | − | − | + | + | + | + | − |
| 2 | Нижние листья отчетливо черешковые | − | − | + | − | − | − | + | − |
| 3 | Семена морщинистые | + | | − | + | + | − | − | + |
| 4 | Прицветники пленчатые, с зеленой жилкой | − | − | − | − | − | − | − | + |
| 5 | Чашелистики с резко заметными жилками | + | − | | − | + | − | | − |

Рис. 3. Определительное дерево для видов рода *Stellaria* — Звездчатка, встречающихся в Новосибирской области: a — *S. hebecalux*, b — *S. holostea*, c — *S. bungeana*, d — *S. crassifolia*, e — *S. graminea*, f — *S. longifolia*, g — *S. media*, h — *S. palustris*.

на персональном компьютере занимают не более часа. Подчеркнем, что время расчета существенно зависит от заданного частичного порядка и может быть существенно меньше.

При числе видов, значительно большем указанной границы, предложенный алгоритм может быть использован для проведения вычислений, если учитывать симметрии заданного частичного порядка.

# Список литературы

[1] Кричевский Р. Е. *Сжатие и поиск информации.* Радио и связь, М., 1989.

[2] Рябко Б. Я., Харитонов А. Ю. Метод построения определительных таблиц, обнаруживающих и исправляющих ошибки. *Изв. СО АН СССР. Сер. биол. наук*, вып. 1, 1982.

[3] Krichevsky R. E., Ryabko B. Ya. Universal Retrieval Trees. *Discrete Appl. Math.*, **12**, 1985, 293–302.

[4] Рябко Б. Я. Кодирование источника с неизвестными, но упорядоченными вероятностями. *Проблемы передачи информации*, **14**, №2, 1979, 71–77.

[5] Ермаков Н. Б., Красников А. А., Федотов А. А., Федотов А. М., Хорев А. Г. *База данных "Флора Новосибирской области"*. `http://www-sbras.nsc.ru/win/elbib/bio/db/`

[6] Blahut R. Computation of Channel Capacity and Rate Distortion Functions. *IEEE Trans.*, **JT-18**, No. 4, 1973, 460–473.

[7] Галлагер Р. *Теория информации и надежная связь*. Советское Радио, М., 1974.

[8] Гольштейн Е. Г. *Теория двойственности в математическом программировании и ее приложения*. Наука, М., 1971.

[9] Флора *Сибири*. Т. 6. Наука, Новосибирск, 1993.

[10] Kuhn H. W., Tucker A. W. Non-linear programming. *Proc. of the Second Berkley Symp. of Math. Statistics and Probability*. Berkley and Los Angeles, Univ. of California Press, 1951, 481–492.

[11] Давыдов Э. Г. *Игры, графы, ресурсы*. Радио и связь, М., 1981.

[12] Topsoe F. Частное сообщение.

*Поступила в редакцию 11 мая 1999 г.,*
*в переработанном виде 16 августа 1999 г.*