ИССЛЕДОВАНИЕ ГРАДИЕНТНОГО МЕТОДА ПОСТРОЕНИЯ ОПРЕДЕЛИТЕЛЬНЫХ ТАБЛИЦ, БЛИЗКИХ К ОПТИМАЛЬНЫМ*

Б. Я. Рябко
Сибирский государственный университет телекоммуникаций и информатики, Новосибирск, Россия
e-mail: ryabko@neic.nsk.su

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

The article considers a simple and fast algorithm constructing taxonomic keys close to optimum, i. e. the gradient method. This algorithm has been numerically investigated when the number of objects does not exceed 100. This algorithm constructs taxonomic keys with sufficiently high quality. This allows to recommend the gradient algorithm for practical use.

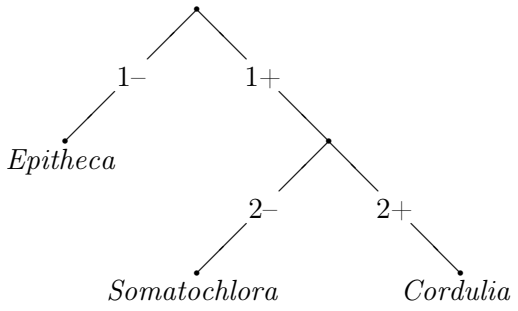
1. Постановка задачи

В биологии определительные таблицы используются для определения таксономической принадлежности видов. Такую таблицу можно представить как двоичное (дихотомическое) дерево, листьям (одновалентным вершинам) которого сопоставлены названия объектов, а развилкам — признаки объектов. Пример ключа из работы [2] показан на рис. 1. Он предназначен для определения родов подсемейства стрекоз Corduliinae.

Определение таксономической принадлежности биологического объекта с помощью ключа можно представить, как движение от корня дерева к одному из листьев. На каждом шаге проверяется признак в текущей развилке, и дальнейшее направление движения выбирается в зависимости от результата проверки. В конце статьи на рис. 3 приведен пример использования определительного ключа для сибирских представителей рода Гвоздика — *Dianthus* из книги [5].

На определение таксономической принадлежности уходит бóльшая часть времени, например, в экологических исследованиях, причем квалифицированно определением могут заниматься только специалисты. Особенно трудоемким является определение насекомых.

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



1–	Окраска тела не зеленая. В основании двух задних крыльев черные пятна.
1+	Окраска тела зеленая. В основании двух задних крыльев нет черных пятен.
2–	Нижние анальные придатки без выроста и короче верхних.
2+	Нижние анальные придатки с выростом и по длине примерно равны верхним.

Рис. 1. Определительная таблица родов подсемейства стрекоз Corduliinae.

При проверке типичных признаков, таких как определение числа жилок на крыльях, приходится использовать микроскопы и бинокуляры.

Таким образом, естественно возникает задача построения ключа, при использовании которого среднее время определения было бы минимально. В этой статье мы опишем простой алгоритм, с помощью которого достаточно хорошо решается поставленная задача.

Для исследования качества такого алгоритма нам необходимо задать множество начальных данных, на которых он будет тестироваться, и критерий качества получаемых определительных таблиц. При оценке качества ключа мы делаем два предположения: представители всех таксонов встречаются с равными вероятностями и сложность проверки всех признаков одинакова.

Поясним, почему была выбрана именно эта модель, несмотря на то, что она весьма условна. Так, трудоемкость верного определения вида сибирских гвоздик *D. borbasii* значительно меньше трудоемкости проверки любого другого признака, потому что этот вид существенно отличается от всех остальных. Вместе с тем, в силу простоты предложенной модели, ее можно рассматривать в качестве удобного “первого приближения” к поставленной задаче. Тем более, что созданный программный инструментарий позволяет провести аналогичное исследование для любой другой модели с заданными значениями частот встречаемости видов и трудоемкостей проверки признаков. Отдельного исследования заслуживает вопрос придания разумных числовых значений этим величинам.

Будем оценивать качество ключа средним временем (трудоемкостью) определения объекта. Если количество типов объектов обозначить через N, то в наших предположениях среднее время определяется формулой $C = (L(1) + L(2) + \cdots + L(N))/N$, где $L(i)$ — длина пути в двоичном дереве от корня до соответствующего листа, равная числу проверяемых признаков. Например, среднее время определения объекта в дереве, изображенном на рис. 1, равно 5/3. Из всех таблиц, построенных по данному набору признаков, таблицу с наименьшим средним временем определения объекта будем называть оптимальной.

Рассмотрим построение определительной таблицы для рода Мшанка — *Sagina*. Исходные данные для этого примера подготовлены по определительной таблице и морфологическим описаниям из книги [5] и содержатся в табл. 1

По этим данным можно построить несколько определительных таблиц. Две из них изображены на рис. 2. Средняя трудоемкость левой таблицы равна 2, в то время как правой — 2.25.

Таблица 1

Признаки для рода Мшанка — *Sagina*: a — *S. intermedia*, b — *S. nodosa*, c — *S. procumbens*, d — *S. saginoides*

№ пп.	Признаки	Таксоны			
		a	b	c	d
1	Количество чашелистиков, лепестков и створок коробочки (– по 5; + по 4)	+	–	+	–
2	Соотношение длины лепестков и чашелистиков (– лепестки вдвое длиннее чашелистиков; + лепестки равны или короче чашелистиков)	+	–	+	+
3	Стебли (– не укореняющиеся; + укореняющиеся)	–	–	+	–

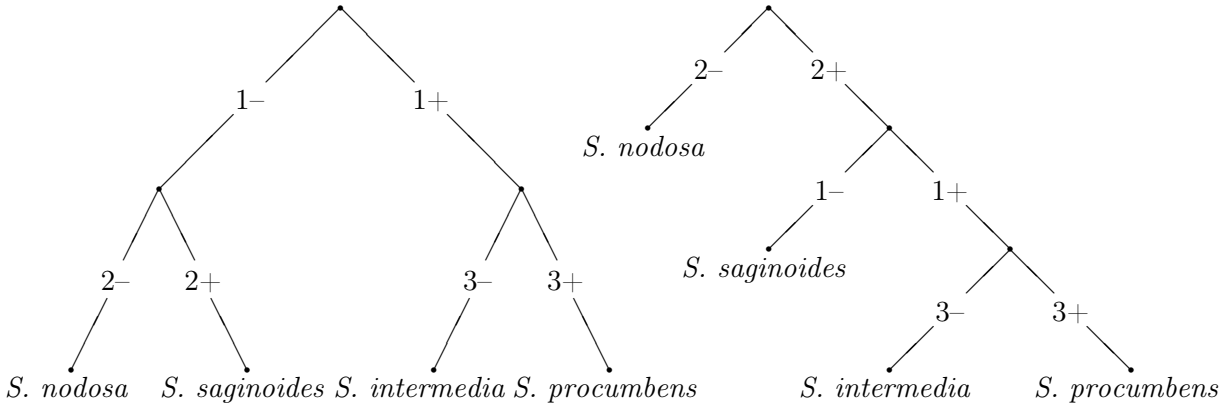


Рис. 2. Определительные деревья для рода Мшанка — *Sagina*

В книге [5] предлагается применять левую часть таблицы. В теории информации [1] доказывается, что средняя трудоемкость определительной таблицы не может быть меньше двоичного логарифма количества объектов. Таким образом, левая таблица является оптимальной.

Известно, что поиск оптимальной определительной таблицы по заданному набору признаков — NP-полная задача [4]. Неформально это означает, что данная задача решается полным перебором всех определительных таблиц и время, необходимое для ее решения, экспоненциально растет с ростом числа признаков n. Таким образом, построение оптимального ключа возможно лишь при небольших n.

Поэтому мы ограничимся поиском таблицы, близкой к оптимальной. Естественный алгоритм решения этой задачи подробно описан во втором разделе. Его работа выглядит следующим образом. Сначала выбирается корневой признак. Он делит множество объектов на две по возможности равные части. Затем для этих меньших частей аналогично строятся свои определительные таблицы.

Так как на каждом шаге выбор корневого признака совершается по возможности наилучшим образом, мы называем такой алгоритм градиентным. Время работы этого алгоритма пропорционально mn, где m — количество подлежащих определению объектов. Таким образом, градиентный алгоритм работает значительно быстрее полного перебора.

Как будет показано в третьем разделе, при числе объектов менее 100 качество деревьев, построенных градиентным алгоритмом с использованием небольшого количества слу-

чайных признаков, в среднем лишь немного уступает качеству оптимальных деревьев, построенных с использованием всех возможных признаков.

2. Градиентный алгоритм

Здесь мы приведем неформальное описание рассматриваемого алгоритма, демонстрируя его работу на примерах. Построим определительную таблицу, исходя из данных, содержащихся в табл. 1

Градиентный алгоритм строит искомое двоичное дерево последовательно, начиная от корня. Очередной развилке приписывается признак, минимизирующий заданную оценочную функцию. Если признак делит все множество объектов на два множества с числом элементов p и q, то в качестве оценочной функции можно взять модуль разности $|p - q|$. Заметим, что оценочную функцию минимизирует признак, который делит все множество объектов приблизительно пополам.

В первом столбце табл. 2 показаны значения оценочной функции. Эти значения — количественная мера того, насколько хорошо каждый из признаков делит множество всех объектов. Минимум оценочной функции достигается на первом признаке.

После того как первый признак будет сопоставлен корню, необходимо построить определительные деревья для образовавшихся частей. В левой развилке объекты *S. nodosa* и *S. saginoides* может разделить только второй признак, а в правой объекты *S. intermedia* и *S. procumbens* — только третий.

Мы получаем определительное дерево, которое совпадает с оптимальным. Как показывает следующий пример, это не всегда так. В табл. 3 содержатся признаки для определения по чашечке таксономической принадлежности гвоздик, встречающихся в Средней Сибири [5].

В этом случае все признаки делят множество из шести объектов на подмножества из четырех и двух объектов. Таким образом, с точки зрения оценочной функции, они неразличимы. Поэтому в качестве корневого градиентный алгоритм может выбрать первый признак. Получающееся при этом дерево изображено на рис. 3. Его средняя трудоемкость равна 17/6. На рис. 4 показано оптимальное дерево, средняя трудоемкость которого равна 16/6.

На рис. 8 в конце статьи приведено описание градиентного алгоритма на формальном алгоритмическом языке. Алгоритм Grad(M, N) получает в качестве входа множества объектов и признаков. Результатом его работы является двоичное дерево, одновалентным вершинам которого сопоставлены названия таксонов, а развилкам — признаки. В описании использованы типы данных дерево и множество, реализация которых описана, например, в [6].

Таблица 2

Значения оценочной функции на признаках из таблицы 1 в трех развилках

Признак	Корень	Левая развилка	Правая развилка
	S. intermedia, S. nodosa, S. procumbens, S. saginoides	*S. nodosa, S. saginoides*	*S. intermedia, S. procumbens*
1	$\|2-2\|=0$	$\|2-0\|=2$	$\|0-2\|=2$
2	$\|1-3\|=2$	$\|1-1\|=0$	$\|0-2\|=2$
3	$\|3-1\|=2$	$\|2-0\|=2$	$\|1-1\|=0$

Т а б л и ц а 3

Признаки для определения таксономической принадлежности гвоздик, встречающихся в Средней Сибири: a — *D. deltoides*, b — *D. ramosissimus*, c — *D. repens*, d — *D. superbus s. str.*, e — *D. superbus* subsp. *sajanensis*, f — *D. versicolor*

№ пп.	Признаки	Таксоны					
		a	b	c	d	e	f
1	Зубцы по краю лепестка — острые	+	−	−	−	−	+
2	Ширина чашечки менее 4 мм	+	+	−	−	−	−
3	Отношение длины чашечки к ширине менее 4	−	−	+	−	−	+
4	Цилиндрическая чашечка	+	−	−	−	+	−

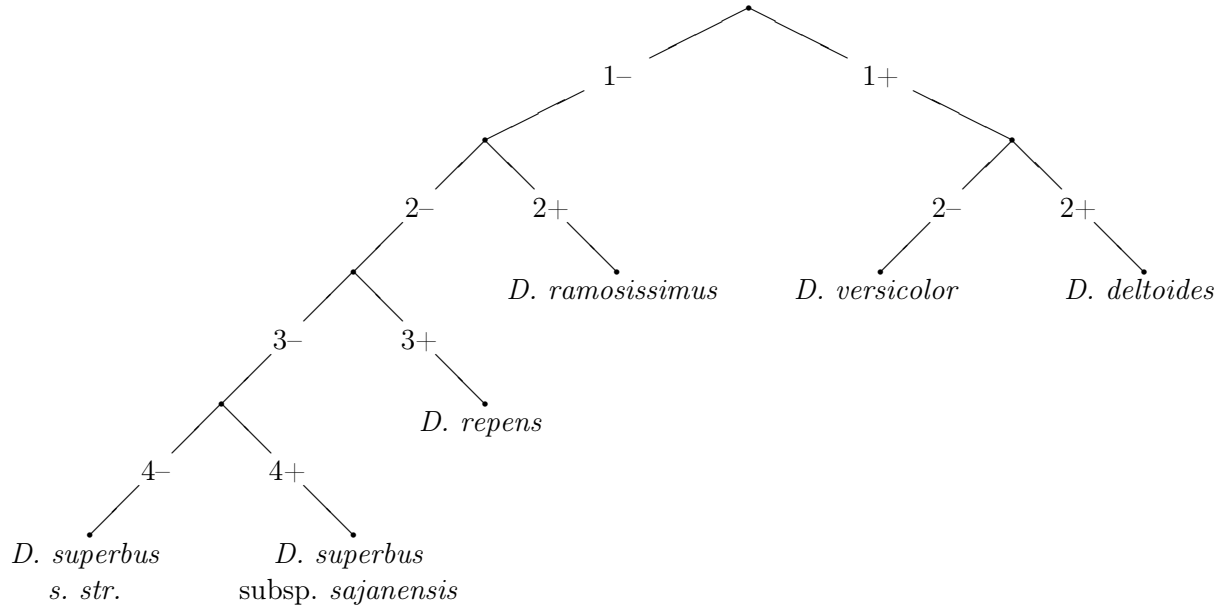


Рис. 3. Определительное дерево для гвоздик Средней Сибири, построенное по градиентному алгоритму.

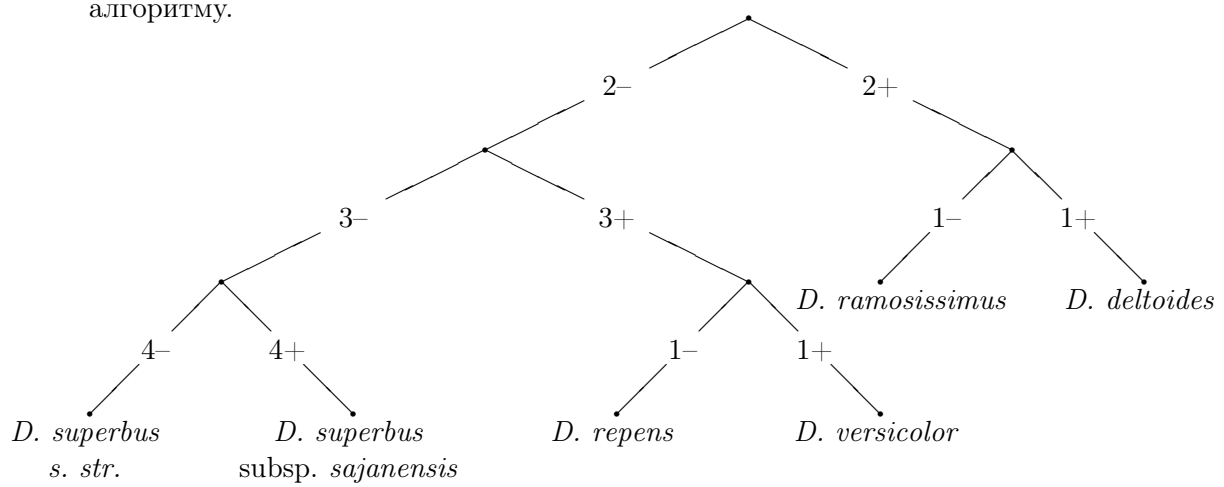


Рис. 4. Оптимальное определительное дерево для гвоздик Средней Сибири.

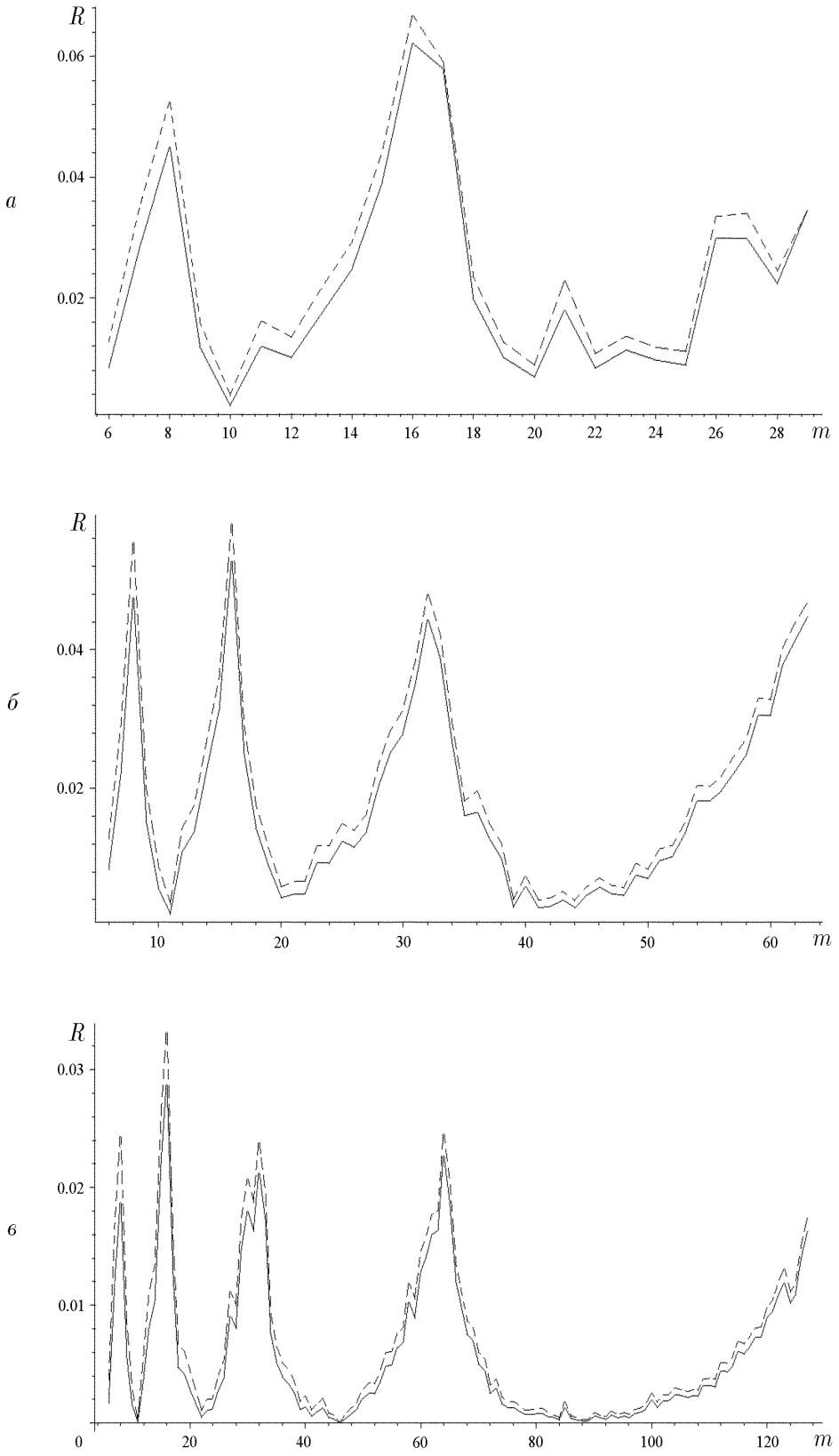


Рис. 5. Выборочное среднее и доверительный интервал для избыточности деревьев, построенных градиентным алгоритмом при количестве признаков $n = 1.2 \log_2 m$ (*а*), $n = 1.5 \log_2 m$ (*б*), $n = 2 \log_2 m$ (*в*).

3. Результаты компьютерного эксперимента

Оптимальное дерево для m объектов, построенное с возможностью использования всех признаков, выглядит следующим образом [1]. В нем k ветвей по длине равны $l = \lfloor \log_2 m \rfloor$, а $m-k$ ветвей на 1 длиннее. Легко видеть, что $k+m = 2^{l+1}$. Таким образом, среднее время поиска C по такому дереву равно $l + 2 - 2^{l+1}/m$. Избыточностью R дерева, построенного с помощью градиентного алгоритма, будем называть разность среднего времени поиска в этом и оптимальном деревьях.

Заметим, что нижняя граница для среднего времени поиска, возможно, меньше среднего времени поиска в оптимальном дереве, построенном с использованием данных признаков. Зато она легко вычислима и не зависит от выбранного набора признаков. При этом, как показывают расчеты, усредненная избыточность R оказывается весьма малой.

На рис. 5 сплошной линией изображен график зависимости усредненной избыточности R от количества объектов m. При каждом m усреднение производилось по 200 случайным порожденным наборам признаков. Пунктирная линия ограничивает сверху интервал для усредненной избыточности с уровнем доверия 0.95. Количество признаков n выбиралось пропорциональным двоичному логарифму количества объектов m, причем вблизи этой естественной нижней границы. Очевидно, при дальнейшем увеличении числа признаков качество работы алгоритма должно становиться лучше.

Доверительный интервал строился исходя из предположения, что измеряемое среднее подчиняется распределению Стьюдента [7]. Верхняя граница интервала полагалась равной сумме выборочного среднего и выборочной дисперсии, умноженной на коэффициент, взятый из таблицы для заданного количества измерений.

Можно заметить, что пики избыточности соответствуют m, равным степеням двойки. За исключением этих случаев практически всегда среднее значение избыточности не превышает 0.05. Таким образом, мы видим, что градиентный алгоритм дает почти оптимальное дерево даже по небольшому количеству признаков, что позволяет рекомендовать его для практического использования.

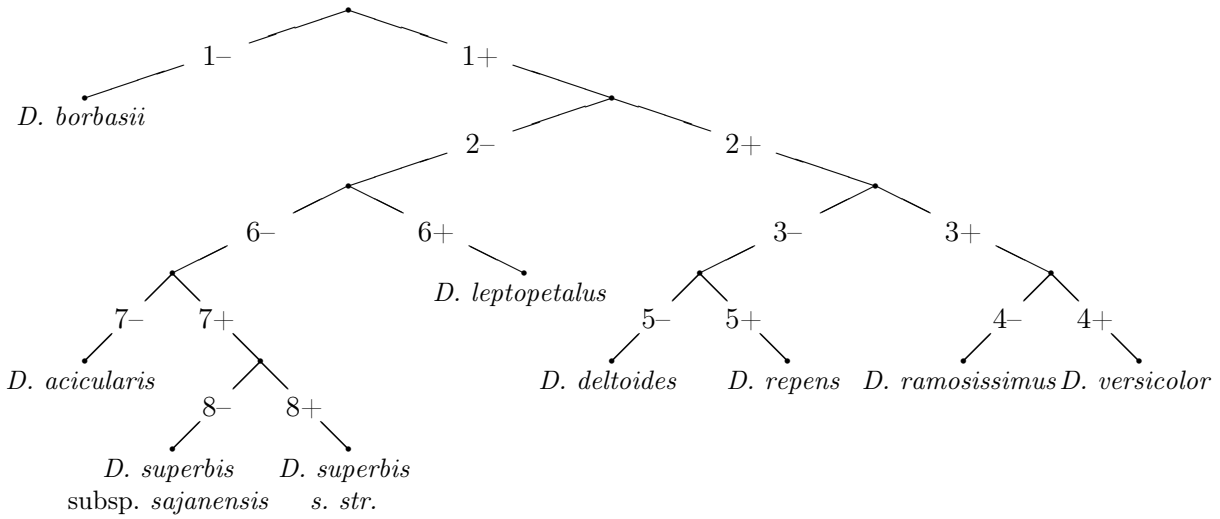
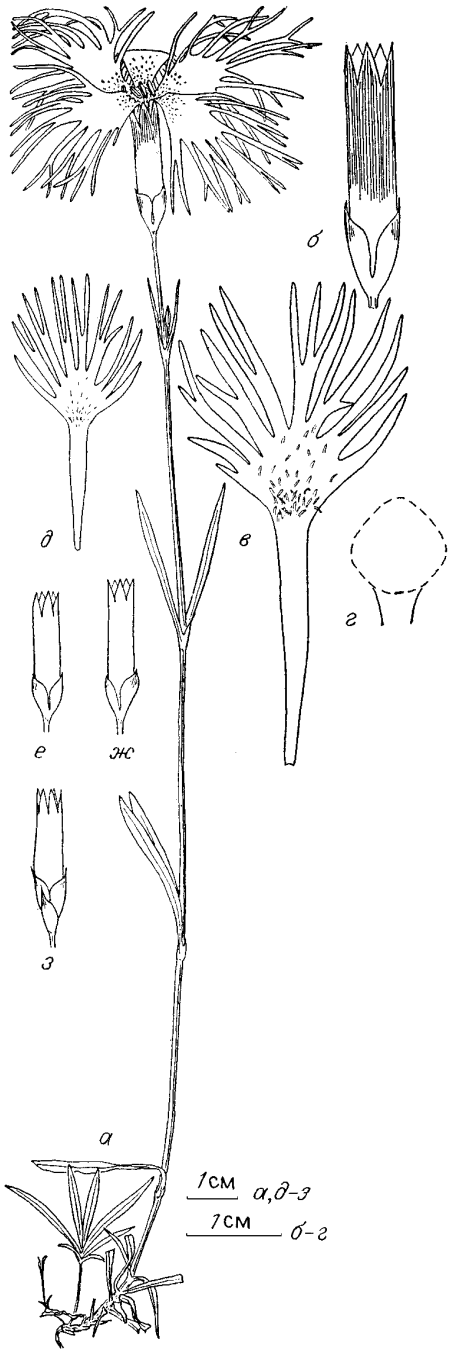


Рис. 6. Определительное дерево рода Гвоздика — *Dianthus*

Таблица 4

Список ступеней ключа (определительных признаков) для сибирских представителей рода *Dianthus* — Гвоздика

1–	Длина листовых влагалищ в 3–5 раз превышает диаметр стебля. Цветки собраны по 2–9 в плотное головчатое верхушечное соцветие с 2 общими присоцветными чешуями.
1+	Длина листовых влагалищ равна диаметру стебля или не более чем в два раза превышает его. Цветки одиночные или образуют рыхлое метельчатое соцветие; иногда сближены по 2 вместе (при этом без общих присоцветных чешуй).
2–	Чашечки 19–30 мм дл. Лепестки бахромчато разделены на узколинейные доли или цельные, но тогда цельнокрайние. Прицветные чешуи в 3–4 раза короче чашечки, редко доходят почти до ее середины.
2+	Чашечки 10–18 мм дл. Лепестки цельные, по верхнему краю мелкозубчатые. Прицветные чешуи равны половине длины чашечки и больше, часто доходят до основания зубцов чашечки и еще выше; иногда короче середины чашечки, тогда лепестки желтовато-белые, 3.5–5.5 мм шириной.
3–	Прицветные чешуи в числе 2 или отсутствуют; если их 4, то наружные обычно более длинные и узкие, листовидные, чашечки часто наверху расширенные, с притупленными треугольными зубцами, длина которых не более чем в два раза превышает их ширину.
3+	Прицветные чешуи в числе 4 или 6(8), б. м. одинаковые по форме; если их иногда 2, то чашечки наверху б. м. суженные, с ланцетными, тонко заостренными зубцами, длина которых в два раза и более превышает их ширину.
4–	Лепестки белые или желтовато-белые, снаружи зеленовато-белые, 3.5–5.5 мм шир. Прицветные чешуи в числе 6, реже 4 или 8, в 2–2.5(3) раза короче чашечки. Листья узколинейные, иногда вдоль сложенные, 1.5–4.5 см дл., 0.8–2 мм шир.
4+	Лепестки пурпуровые или розовые, снаружи зеленовато-розовые, 7–12 мм шир. Прицветные чешуи обычно в числе 4, реже 6, доходят до середины чашечки и выше. Листья линейно-ланцетные или линейные, 2–7 см дл., 1.5–4 мм шир.
5–	Растения коротко бархатисто опушенные. Корневище тонкое, ползучее, 1–2 мм диам. Стебли малочисленные, в верхней части разветвленные или простые, 1–5-цветковые, 15–35 см выс., с 7–8 междоузлиями. Чашечки относительно узкие, 13–17 мм дл., 2.5–3.5 мм шир. Лепестки 4–7 мм шир.
5+	Растения голые. Корень вертикальный, на верхушке ветвистый, 2–5 мм диам. Стебли многочисленные, простые, одноцветковые, очень редко с единичными укороченными тонкими веточками с недоразвитыми цветками, 6–20 см выс., с 3–5 междоузлиями. Чашечки 9–14 мм дл., 4.5–8 мм шир. Лепестки 8–12(15) мм шир.
6–	Листья плоские, линейно-ланцетные или линейные; если игловидные, то короткие (1–3 см дл.), а чашечки 20–24 мм дл. Венчик пурпуровый, розовый или белый. Лепестки 7–25 мм шир., с волосками у основания отгиба, разделены бахромчато на узколинейные доли.
6+	Листья щетиновидные, вдоль свернутые, жесткие, 2–10 см дл. Чашечки 25–30 мм дл. Венчик желтовато-зеленоватый, снаружи с коричневым оттенком. Лепестки 3–6 мм шир., без волосков у основания отгиба, цельные, цельнокрайние.
7–	Листья игловидные, килеватые, 1–3 см дл., 0.5–1.2 мм шир. Отгиб лепестков обратнояйцевидный, 12–15 мм дл., 7–10 мм шир., примерно на 1/3 надрезанный на многочисленные линейные лопасти.
7+	Листья линейно-ланцетные или линейные, плоские, 3–10 см дл., 2–8 мм шир. Отгиб лепестков широкоэллиптический или обратноширокояйцевидный, 17–25(30) мм дл., ок. 20 мм шир., на 1/2 и глубже разделенный на многочисленные узколинейные доли, которые могут быть еще разделены.
8–	Прицветные чешуи в числе 2–4, пурпурово-фиолетовые или фиолетово-черные, часто с сизым налетом. Чашечки наверху обычно не суженные, 4.5–6 мм шир. Линейные доли лепестков по длине примерно равны ширине нерассеченной центральной части, которая округло-ромбовидная или широкояйцевидная, 8–11 мм дл. и шир.
8+	Прицветные чешуи в числе 4, редко 2 или 6, коротко заостренные, зеленые или фиолетово окрашенные. Чашечки наверху немного суженные, 3.5–5(5.5) мм шир. Линейные доли лепестков в 2–3(4) раза длиннее нерассеченной центральной части, которая продолговато-эллиптическая, продолговато-яйцевидная до яйцевидной, редко обратнояйцевидная, 6–13 мм дл., 3–7 мм шир.



Пример использования определительного ключа

Слева приведен рисунок из книги [5], на котором изображена гвоздика (*а* — внешний вид, *б* — чашечка, *в, д* — лепесток, *г* — нерассеченная часть лепестка, *е–з* — чашечка с прицветными чешуями).

Определим ее таксономическую принадлежность с помощью определительного ключа из той же книги (см. рис. 6, табл. 4).

Так как цветки одиночные, выполнен признак 1+. Переходим к проверке второго признака. Чашечка на рисунке несколько длиннее 2 см, лепестки бахромчато разделены на узколинейные доли, прицветные чешуи в 3 раза короче чашечки — выполнен признак 2–. Переходим к проверке шестого признака. Лепестки бахромчато разделены на узколинейные доли. Выполнен признак 6–, проверим признак 7. Листья плоские. Лепестки на 1/2 разделены на многочисленные узколинейные доли. Выполнен признак 7+. Проверим признак 8. Чашечка наверху не суженная, приблизительно 5 мм шириной. Линейные доли лепестков приблизительно равны нерассеченной центральной части. Итак, выполнен признак 8–.

Таким образом, можно сделать заключение, что изображенная на рисунке слева гвоздика принадлежит к подвиду *Dianthus superbus* subsp. *sajanensis* (гвоздика саянская). Заметим, что определение реальных биологических видов значительно сложнее в связи с большой изменчивостью их представителей. Более того, данное обстоятельство вынуждает часто обновлять определительные таблицы.

Рис. 7. Определение таксономической принадлежности с помощью ключа.

```
алг Оценка(признак p, множ M): число;
нач
  Оценка := Card(p.лев(M))*Card(p.прав(M))
кон;

алг Grad(множ M, N): дерево;
  число t;
  признак q;
нач
  если Card(M) = 1 то
    Grad.тип := лист;
    Grad.таксон := элемент(M)
  иначе
    Grad.тип := развилка;
    t := 0;
    для p из N
    цикл
      если Оценка(p, M)>t то
        t := Оценка(p, M); q := p
      ку
    кц;
    Grad.признак := q;
    Grad.лев := Grad(q.лев(M), N);
    Grad.прав := Grad(q.прав(M), N)
  ку
кон;
```

Рис. 8. Градиентный алгоритм.

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, 124–130.

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

[4] Гэри М., Джонсон Д. *Вычислительные машины и труднорешаемые задачи.* Мир, М., 1982.

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

[6] Вирт Н. *Алгоритмы + структуры данных = программы.* Мир, М., 1985.

[7] Кендалл Н. Дж., Стюарт А. *Статистические выводы и связи.* Наука, М., 1973.

Поступила в редакцию 28 января 1999 г.,
в переработанном виде 16 марта 1999 г.