Ученые записки Таврического национального университета им. В. И. Вернадского

Серия "ИНФОРМАТИКА И КИБЕРНЕТИКА" № 1 (2001) 36 – 38

УДК 519.8

АСИМПТОТИКА ЧИСЛА БИНАРНЫХ РЕШАЮЩИХ ДЕРЕВЬЕВ

Донской В. И.

Бинарные решающие деревья (БРД) являются алгоритмическими соответствиями, предназначенными для вычисления функций вида $f\colon \{0,1\}^n \to \{0,1,\ldots,(k-1)\}$. Они широко применяются в информатике для решения задач распознавания образов, формирования понятий, оптимизации при неполных данных [1,2]. Класс функций, вычисляемых при помощи БРД, обозначим P_{2-k}. Функции из P_{2-k} принимают значения $\{0,1,\ldots,k-1\}$, а их аргументы — булевы. Каждое БРД T вычисляет ровно одну функцию $f_T \in P_{2-k}$. Обратное неверно: произвольная функция $f \in P_{2-k}$ может быть вычислена, вообще говоря, более чем одним БРД. Поэтому мощность класса P_{2-k} меньше мощности класса $P_{\text{БРД}}$ всех БРД с булевыми переменными, используемыми во внутренних вершинах, и k-значными метками листьев [1].

Обозначим $D(n,k,\mu)$ класс БРД, в котором каждое дерево имеет ровно μ листьев; n — число булевых переменных; k — число возможных пометок листьев. Число различных БРД $d(n,k,\mu) = |D(n,k,\mu)|$ представляет интерес в связи с его использованием для оценки статистической надежности решающих правил, основанных на построении деревьев, и изучения алгоритмов синтеза БРД с наименьшим числом листьев [1,3].

Точная формула для $d(k,n,\mu)$ неизвестна. Ниже получена асимптотика для $d(k,n,\mu)$.

Теорема. *Мощность класса $D(n,k,\mu)$ при заданных константах k и μ имеет асимптотику*

$$d(n,k,\mu) \sim (\mu-1)!\,[k(k-1)]^{\mu-1} n(n-1)^{\mu-2} \quad \textit{при} \quad n \to \infty.$$

Доказательство. Очевидно, что $d(n,k,2) = n \cdot 2 \cdot C_k^2$, где C_k^2 — число сочетаний из k по два. Переход к БРД с $j+1$ листьями связан с заменой в некотором БРД с j листьями одной концевой вершины на новую внутреннюю и добавлением двух новых листьев. Такой процесс "достройки" предполагает:

а) выбрать любой из j листьев;

б) заменить выбранный лист вершиной с переменной, не встречавшейся в ветви, которая заканчивалась заменяемым листом;

в) выбрать два значения из $\{0,1,\ldots,(k-1)\}$ для пометки двух новых листьев.

Поскольку наибольшее возможное число неконцевых вершин в ветви БРД с j листьями равно $j-1$, то выбрать переменную для новой внутренней вершины

можно не менее (причем для некоторых ветвей — строго менее) чем $n-j+1$ способами. При $n \geqslant j$ очевидно неравенство:

$$d(n,k,j+1) > d(n,k,j) \cdot j \cdot 2 \cdot C_k^2 \cdot (n-j+1). \tag{1}$$

Из неравенства (1) получается нижняя оценка:

$$d(n,k,\mu) > L(n,k,\mu) = n(\mu-1)!\,[k(k-1)]^{\mu-1}(n-\mu+2)^{\mu-2}.$$

С другой стороны, выбрать переменную для замены листа внутренней вершиной можно, вообще говоря, менее чем $n-1$ способами, поэтому

$$d(n,k,j+1) < d(n,k,j) \cdot j \cdot 2 \cdot C_k^2 \cdot (n-1). \tag{2}$$

Из неравенства (2) получается верхняя оценка:

$$d(n,k,\mu) < H(n,k,\mu) = n(\mu-1)!\,[k(k-1)]^{\mu-1}(n-1)^{\mu-2}.$$

Легко убедиться, что

$$\lim_{n\to\infty} \frac{L(n,k,\mu)}{H(n,k,\mu)} = 1.$$

Следовательно, $H(n,k,\mu) \sim L(n,k,\mu)$ при $n \to \infty$ и $d(n,k,\mu) \sim (\mu-1)!\,[k(k-1)]^{\mu-1} n(n-1)^{\mu-2}$. □

Следствие 1. *Число булевых функций от n переменных $b(n,2,\mu)$, представимых БРД с ровно μ листьями, удовлетворяет неравенству*

$$b(n,2,\mu) < (\mu-1)!\,2^{\mu-1} n^{\mu-1}.$$

Следствие 2. *Класс $P^2_{\text{БРД}}(n,\mu)$ булевых функций, представимых БРД с ровно μ листьями при $n \to \infty$ сколь угодно узок по сравнению с классом $L(n)$ линейных функций из $P_2(n)$.*

Доказательство.

$$|L(n)| = 2^{n+1}; \qquad \frac{|P^2_{\text{БРД}}(n,\mu)|}{|L(n)|} < \frac{b(n,2,\mu)}{|L(n)|};$$

$$b(n,2,\mu) = o(2^{n+1}), \quad n \to \infty.$$

□

Следствие 3. *Для любой заданной константы μ при $n \to \infty$ задача вычисления свойства "Существует БРД с не более чем μ листьями, корректное на непротиворечивой обучающей выборке длины m_0" полиномиально разрешима.*

Доказательство следует из существования переборного алгоритма синтеза всех БРД, начиная с деревьв класса $D(n,k,2)$, до всех деревьев класса $D(n,k,\mu)$. В процессе такого синтеза будет выполнено не более μ этапов построения, на которых для не более чем $c_1 n^{\mu-2}$ деревьев будет реализовано не более $c_2 \mu m_0 n$ проверок корректности и не более $c_3 n^{\mu-1}$ "достроек", где c_1, c_2, c_3 — константы.

Замечание. Следствие 3 имеет скорее теоретическое значение, поскольку при числе переменных n порядка сотен и величине константы μ порядка нескольких десятков полиномиальная оценка трудоемкости вида $O(n^{\mu})$ не позволяет надеяться на построение быстрого алгоритма синтеза корректного БРД с наименьшим числом листьев μ. Тем не менее, синтез приближенного (некорректного) БРД с μ^* листьми, обеспечивающего минимальное число неправильно классифицируемых точек из обучающей выборки, можно осуществить перебором. Константа μ^* может быть выбрана из статистических соображений [1].

Список литературы

1. Донской В. И. *Алгоритмы обучения, основанные на построении решающих деревьев* // ЖВМ и МФ, 1982, 22, 4, с.963–974.
2. Donskoy V. *Pseudo-Boolean Scalar Optimization Models with Incomplete Information* // GMOOR Newsletter, 1996, 1/2, p. 20–26.
3. Вапник В. Н. *Восстановление зависимостей по эмпирическим данным*. М.: Наука, 1979. 448 с.