Вики

Алгоритм ID3 (ID3 algorithm)

> Синонимы: Iterative Dichotomizer-3
> Разделы: Алгоритмы

В анализе данных и машинном обучении ID3 — это один из наиболее популярных алгоритмов обучения деревьев решений. В основе идеи алгоритма лежит рекурсивное разбиение обучающего множества, размещаемого в корневом узле дерева решений, на подмножества с помощью решающих правил.

Разбиение продолжается до тех пор, пока в результирующих подмножествах не останутся примеры только одного класса, после чего процесс обучения остановится, а подмножества будут объявлены листами дерева, содержащими решения.

Каждый атрибут обучающего множества отражает некоторое свойство классифицируемых объектов. При этом атрибуты могут иметь разную значимость с точки зрения классификации. Например, атрибут, все значения которого одинаковы, вообще бесполезен для различия классов.

Классифицирующая сила других атрибутов может быть разной. Целью алгоритма является выбор атрибутов для разбиения таким образом, чтобы полученное дерево было компактным, простым для понимания и при этом достаточно точным.

Алгоритм начинает работу с корневого узла дерева, который содержит все примеры обучающего множества. На каждой итерации алгоритма выбирается один из атрибутов, по которому производится разбиение множества примеров в узле на подмножества. При этом для дискретных и непрерывных атрибутов процесс отличается.

Дискретный атрибут

Пусть атрибут X принимает три значения: A, B и C. Тогда при разбиении исходного множества T по атрибуту X алгоритм сформирует три узла-потомка $T_1(A)$, $T_2(B)$ и $T_3(C)$, в первом из которых будут содержаться все примеры, в которых атрибут X принимает значение A, во втором — значение B, и в третьем — C. Процесс рекурсивно повторяется до тех пор, пока не будут сформированы подмножества, содержащие примеры только одного класса.

Выбор атрибута на каждом разбиении производится с помощью критерия прироста информации:

$$Gain(X) = Info(T) - Info_X(T),$$

где $Info(T)$ — информация множества до разбиения, $Info_X(T)$ — информация после разбиения по атрибуту X. В качестве атрибута разбиения выбирается атрибут, который обеспечивает максимальное значение $Gain(X)$.

Непрерывный атрибут

Если атрибут, по которому производится разбиение, непрерывный, то его сначала преобразуют в дискретный вид, например, с помощью операции квантования. Затем значения ранжируются и ищется среднее, которое используется для выбора порога. Все примеры, имеющие значения атрибута выше порогового, помещаются в один узел-потомок, а те, которые ниже — в другой.

Таким образом, при использовании непрерывного атрибута узлы дерева имеют по два потомка.

Недостатком алгоритма является склонность к переобучению, особенно в случае, когда разнообразие значений атрибута велико. В пределе, если все значения атрибута уникальны, мы можем получить дерево с числом узлов, равным числу примеров. Кроме этого, алгоритм не предусматривает возможности работы с пропусками в данных.

Алгоритм разработан Джоном Р. Квинланом в 1986 году. Позднее алгоритм был доработан, и новая версия получила название C4.5. В ней была решена проблема переобучения и стала доступной обработка пропусков в данных.