Вики

Алгоритм CART (CART algorithm)

> Синонимы: Classification and Regression Tree
> Разделы: Алгоритмы

Популярный алгоритм построения деревьев решений, который может работать как с дискретной, так и с непрерывной выходной переменной, т.е. решать задачи и классификации, и регрессии.

Алгоритм строит бинарные деревья решений, которые содержат только два потомка в каждом узле. В процессе работы происходит рекурсивное разбиение примеров обучающего множества на подмножества, записи в которых имеют одинаковые значения целевой переменной.

Алгоритм реализует обучение с учителем и использует в качестве критерия для выбора разбиений в узлах индекс чистоты Джини (Gini impurity index). В процессе роста дерева алгоритм CART проводит для каждого узла полный перебор всех атрибутов, на основе которых может быть построено разбиение, и выбирает тот, который максимизирует значение индекса Джини.

Основная идея алгоритма заключается в том, чтобы выбрать такое разбиение из всех возможных в данном узле, чтобы полученные дочерние узлы были максимально однородными. При этом каждое разбиение производится только по одному атрибуту.

Если атрибут X, по которому производится разбиение, является номинальным с I категориями, то для него существует $2^{(I-1)}$ возможных разбиения, а если порядковым или непрерывным с K различными значениями, существует $K-1$ различных разбиений по X. Дерево строится, начиная с корневого узла, путем итеративного использования следующих шагов в каждом узле:

1. Для каждого атрибута ищется лучшее разбиение (в смысле однородности результирующих подмножеств).
2. Среди всех разбиений, найденных на предыдущем шаге, выбирается то, для которого критерий разбиения наибольший.
3. Узел разбивается с использованием лучшего разбиения, найденного на шаге 2, если не выполнено условие остановки.

Процедура упрощения деревьев решений, построенных на основе алгоритма CART, реализуется с помощью специального метода соотношения издержки-сложность (Cost-Complexity Pruning) и перекрестной проверки (для малых наборов данных, где полноценное разделение на обучающее и тестовое множества проблематично).

Алгоритм обладает следующими преимуществами:

- не является статистическим, поэтому не требует вычисления параметров вероятностных распределений;
- атрибуты разбиения выбираются непосредственно в процессе построения дерева, поэтому нет необходимости проводить процедуру отбора переменных для модели;
- устойчив к выбросам и аномальным значениям;
- высокая скорость работы.

К недостаткам алгоритма можно отнести неустойчивость относительно данных: даже небольшие изменения в обучающем множестве порождают значительные изменения в структуре дерева решений.

Алгоритм предложен в 1984 г. Лео Брейманом, Джеромом Фридманом, Ричардом Олшеном и Чарльзом Стоуном.