Алгоритм C4.5 (Algorithm C4.5)

Разделы: Алгоритмы

Алгоритм построения деревьев решений, являющийся улучшенной модификацией алгоритма ID3. Так же, как и ID3, алгоритм C4.5 использует критерий информационной энтропии, или прироста информации (information gain). Обучающий набор данных для алгоритма C4.5 представляет собой множество примеров $S = S_1, S_2, \ldots, S_n$, для которых предварительно задана метка класса. Каждый пример представляет собой p-мерный вектор значений атрибутов $x = x_1, x_2, \ldots, x_p$.

В каждом узле дерева решений производится выбор атрибута, который позволяет разбить множество попавших в него примеров на подмножества, максимально «чистые» по классовому составу. Чем однороднее полученные подмножества, тем больше их информация и меньше энтропия.

Следовательно, атрибут, выбираемый для разбиения в узле, должен обеспечивать максимальный прирост информации, или уменьшение энтропии, в результирующих подмножествах. Алгоритм рекурсивно продолжает процедуру разбиения до тех пор, пока в листьях не останутся примеры одного класса. Затем производится упрощение дерева решений путем отсечения ветвей.

Алгоритм C4.5 имеет следующие отличия от ID3:

- возможность работы как с непрерывными, так и с дискретными атрибутами;
- возможность обучения на данных, содержащих пропуски (пропущенные значения помечаются и при вычислениях не используются);
- использует для упрощения дерева решений метод отсечения ветвей, когда сначала строится полное дерево, а затем его размер сокращается путем преобразования узлов в листья.

Алгоритм C4.5 разработан Дж. Р. Куинленом (John Ross Quinlan), который является также автором алгоритма ID3.