Критерий прироста информации (Information Gain)

Разделы: Метрики

В анализе данных и машинном обучении критерий прироста информации — это критерий, используемый для выбора лучшего разбиения подмножеств в узлах деревьев решений в алгоритмах обучения ID3 и C4.5.

В процессе обучения деревьев решений производится рекурсивное разбиение узлов на узлы-потомки, которые должны быть более однородными по классовому составу попавших в них примеров, чем родительский узел. Следовательно, энтропия дочерних узлов должна быть меньше, чем родительских, а внутренняя информация — больше.

Разбиение в каждом узле дерева производится по определенному атрибуту из обучающего множества. Поскольку атрибутов несколько, при каждом разбиении приходится решать задачу выбора наилучшего атрибута. Наилучшим атрибутом будет считаться тот, который обеспечит максимально возможное увеличение однородности классов в дочернем узле относительно родительского или, что одно и то же, максимальное снижение энтропии (прирост информации).

Таким образом, критерий прироста информации реализует следующую последовательность действий:

1. Строятся разбиения по всем доступным атрибутам.
2. Вычисляется прирост информации (уменьшение энтропии) узла T в результате разбиения по атрибуту A. $Gain(A) = Info(T) - Info_A(T)$.
3. Выбирается атрибут, разбиение по которому обеспечит наибольший прирост информации.

В большинстве случаев применение критерия прироста информации для определения значимости атрибутов показывает хорошие результаты. Проблемы возникают, когда атрибут имеет большое разнообразие уникальных значений. В этом случае дерево решений оказывается склонным к переобучению.

Для решения данной проблемы в алгоритмах C4.5 и C5.0 вместо критерия прироста информации используется отношение прироста информации (gain-ratio).