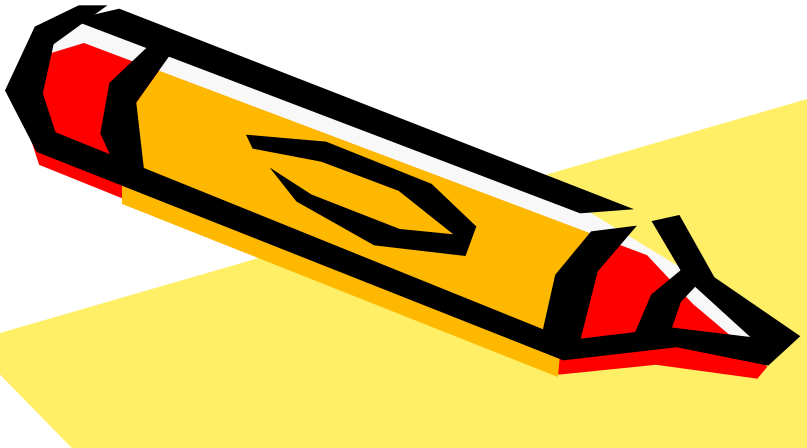


Системы искусственного интеллекта

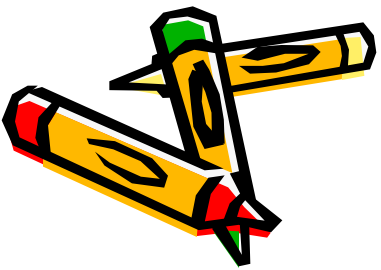
Бондаренко Иван Юрьевич,
ассистент каф. ПМИ

Алгоритм C4.5

Усовершенствование простого алгоритма индуцирования знаний.

Основное отличие: следующий условный атрибут, по которому проводится разбиение, определяется по критерию минимизации энтропии.

Теперь алгоритм не зависит от порядка следования атрибутов таблицы данных.



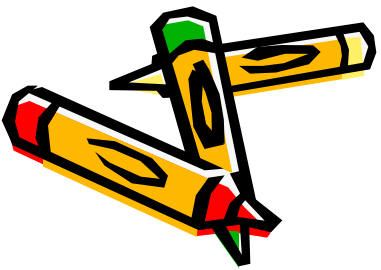
Общее описание алгоритма C4.5

Алгоритм работает для таких таблиц данных, в которых атрибут класса (целевой атрибут) может иметь конечное множество значений.

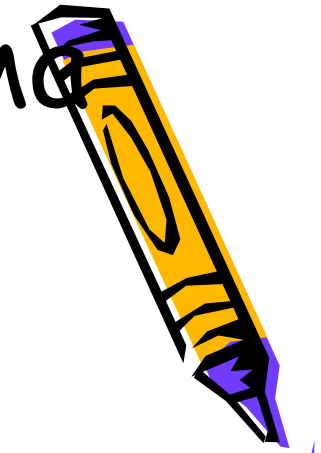
Обозначения

T — множество примеров (таблица или подтаблица данных);

m — количество условных атрибутов (столбцов таблицы)



Общее описание алгоритма C4.5 (продолжение)



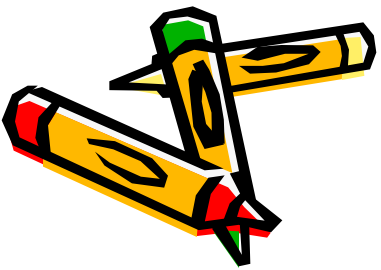
Обозначения (продолжение)

$|T|$ — мощность множества примеров (количество строк в таблице или подтаблице данных);

$C_1, C_2, \ldots, C_k$ — значения, принимаемые атрибутом класса;

X — текущий условный атрибут, по которому мы хотим провести разбиение

Общее описание алгоритма C4.5 (окончание)



Обозначения (окончание)

A_1 , A_2 , …, A_N — значения, принимаемые текущим условным атрибутом;

Выбор условного атрибута для разбиения

Пусть рассматриваем условный атрибут X, принимающий n значений $A_1, A_2 \ldots A_n$. Тогда разбиение множества (таблицы) T по атрибуту X даст нам подмножества (подтаблицы) $T_1, T_2 \ldots T_n$.

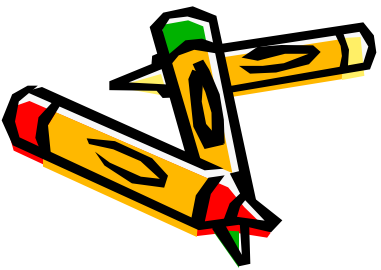
Пусть $freq(C_j, T)$ — количество примеров из множества T, в которых атрибут класса равен C_j

Выбор условного атрибута для разбиения (продолжение)

Тогда вероятность того, что случайно выбранная строка из таблицы T будет принадлежать классу C_j, равна

$$P = \frac{\text{freq}(C_j, T)}{|T|}$$

Например, вероятность того, что прибыль будет расти, составляет $P = 5 / 10 = 0{,}5$



Выбор условного атрибута для разбиения (продолжение)

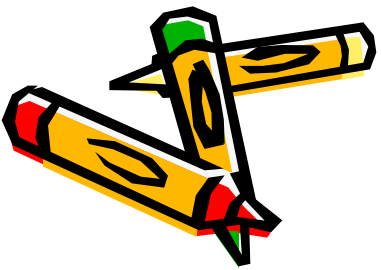
Согласно теории информации, количество содержащейся в сообщении информации зависит от её вероятности $\log_2(1/P)$.

Количество информации измеряется в битах.

Выбор условного атрибута для разбиения (продолжение)

Энтропия таблицы T, то есть среднее количество информации, необходимое для определения класса, к которому относится строка из таблицы T:

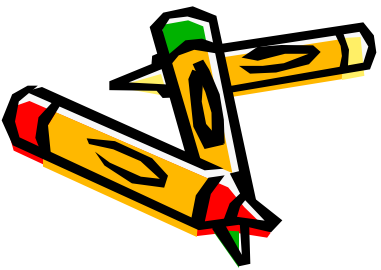
$$\mathrm{Info}(T) = -\sum_{j=1}^{k} \frac{\mathrm{freq}(C_j, T)}{|T|} \cdot \log_2 \left(\frac{\mathrm{freq}(C_j, T)}{|T|} \right)$$



Выбор условного атрибута для разбиения (продолжение)

Энтропия таблицы T после её разбиения по атрибуту X на n подтаблиц:

$$\mathrm{Info}_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \cdot \mathrm{Info}(T)$$



Выбор условного атрибута для разбиения (окончание)

Критерий для выбора атрибута X – следующего атрибута для разбиения:

$$\text{Gain}(X) = \text{Info}(T) - \text{Info}_X(T)$$

Шаги алгоритма C4.5

Шаг 1. Для всех условных атрибутов X_1, … X_m таблицы T вычисляем критерий разбиения $Gain(X_i)$. Выбираем такой атрибут X_i для которого $Gain(X_i)$ максимально.

Шаг 2. Разбиваем таблицу по выбран-ному атрибуту на N подтаблиц. Проверяем каждую подтаблицу следующим образом.

2.1. Если подтаблица монотонна (все строки относятся к одному классу), то порождаем правило.

2.2. В противном случае рекурсивно применяем алгоритм C4.5 к полученной подтаблице