**Difference between ID3 and C4.5**

The definition of the gain used in C4.5 is different from that used in ID3:

In ID3, Gain is defined as

Gain(A, T) = Info(T) – Info(A, T)

The C4.5, Gain is defined as

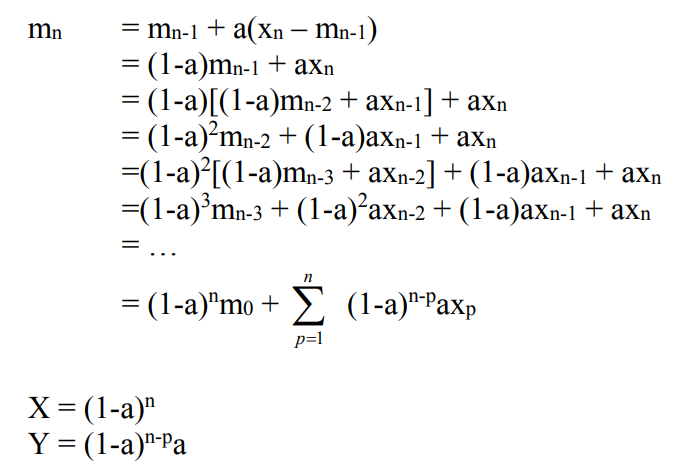


where .

In ID3, there is a higher tendency to choose an attribute containing more distinct values. Therefore, to prevent this, the SplitInfo factor in C4.5 is used to penalize an attribute that containing more distinct values. If this more distinct values, the greater the penalty.

**Forgetful k-means Mining**

* Make initial guesses for the means m1, m2, …, mk
* Until Interrupted
  + Acquire the next example x
  + If mi is closest to x,
    - replace mi by mi + a(x – mi)



**Sequential k-means Mining**

Make initial guesses of the means m1, m2, …, mk

Set the counts n1, n2, …, nk to zero

Until interrupted

Acquire the next example x

If mi is closest to x

Increment ni

Replace mi by mi + 1/ni (x – mi)

