Assign-1 - Prob-1 - Huffman Encoding Correctness

CSc-59866 - Senior Design - Prof. Wei
Oct 21 2016
Team 2: Jesse Feinman, James Kasakyan,
Ian S. McBride, Jeff Stolzenberg

Source: https://www.cs.princeton.edu/~wayne/kleinberg-tardos/pearson/04Huffman.pdf

Proof by induction for an alphabet S that the huffman code for S is optimal.

Let ABL(c) be defined as follows:

Optimal Prefix Codes

Definition. The average bits per letter of a prefix code c is the sum over all symbols of its frequency times the number of bits of its encoding:

$$ABL(c) = \sum_{x \in S} f_x \cdot |c(x)|$$

Then if the huffman code for S is optimal, it will minimize ABL(c).

For any alphabet S with more than one symbol, the huffman tree T for S contains two elements (y and z) with the lowest frequency at the lowest level.

Observe that:

Claim. $ABL(T') = ABL(T) - f_\omega$

Pf.

$$
\begin{aligned}
\mathrm{ABL}(T) &= \sum_{x \in S} f_x \cdot \mathrm{depth}_T(x) \\
&= f_y \cdot \mathrm{depth}_T(y) + f_z \cdot \mathrm{depth}_T(z) + \sum_{x \in S, x \neq y,z} f_x \cdot \mathrm{depth}_T(x) \\
&= (f_y + f_z) \cdot (1 + \mathrm{depth}_T(\omega)) + \sum_{x \in S, x \neq y,z} f_x \cdot \mathrm{depth}_T(x) \\
&= f_\omega \cdot (1 + \mathrm{depth}_T(\omega)) + \sum_{x \in S, x \neq y,z} f_x \cdot \mathrm{depth}_T(x) \\
&= f_\omega + \sum_{x \in S'} f_x \cdot \mathrm{depth}_{T'}(x) \\
&= f_\omega + \mathrm{ABL}(T')
\end{aligned}
$$

Proof by induction that Huffman code for S achieves the minimal (optimal) ABL:

Base case: For n = 2, the shortest possible encoding is 1, which is achieved by huffman tree T for S.

Hypothesis: Assume the Huffman tree T* for S* (the alphabet S where the two least frequent symbols y and z are replaced by w, where f(w) = f(y) + f(z)) is optimal

Induction step:

By contradiction:

Suppose Huffman tree T for S is not optimal

Then there exists an optimal tree Z and ABL(Z) < ABL(T)

Let Z* be the tree Z with the two smallest elements (y and z) removed and replaced with w, where f(w) = f(y) + f(z).

From above:

ABL(Z*) = ABL(Z) - f(w)

ABL(T*) = ABL(T) - f(w)

Since ABL(Z) < ABL(T), then ABL(Z*) < ABL(T*)

But by induction hypothesis, ABL(T*) is optimal. Therefore there cannot exist a tree Z* formed from Z where ABL(Z*) < ABL(T*).

By contradiction, ABL(T) must be optimal.