Laboratorio di Teoria dell'Informazione 2012/13

Stima dell'entropia

L'entropia rappresenta il numero minimo di bit necessari per rappresentare un simbolo di una sorgente di informazione.

Ai fini di questo laboratorio considereremo tre tipi di sorgente di informazione:

1. file di testo in linguaggio naturale, linguaggio di programmazione o misto (ad esempio documento di testo, pagina html, codice sorgente)
2. file multimediale non compresso (ad esempio file audio da CD .wav)
3. file multimediale già compresso (ad esempio jpeg, avi, mp3).

L'entropia di una sorgente senza memoria con probabilità di ogni simbolo p_i è definita come:

$$H = -\sum_{i=0}^{N-1} p_i \log_2 p_i$$

dove *N* rappresenta il numero di simboli.

L'obiettivo di questo laboratorio è quello di stimare le probabilità di ogni simbolo del file, e da queste valutare l'entropia. Infine, tale entropia dovrà essere confrontata con la prestazione di compressione ottenuta con un codificatore a dizionario (zip, gzip, ecc.).

A tale scopo occorre definire il significato di simbolo (o lettera) della sorgente senza memoria nel caso specifico del file. Si definisca con p_i^m la probabilita' del simbolo *i*-esimo del file, ove come simbolo si assume una stringa di *m* bit. Tali probabilità si possono stimare per via numerica valutando le corrispondenti frequenze statistiche. A partire da tali probabilità stimate si può calcolare la entropia di ordine *m*:

$H^{(m)} = -\sum_{i=0}^{2^m-1} p_i^m \log_2 p_i^m$ [bit per stringa di m bit]

ovvero

$H = -H^{(m)}/m$ [bit per bit di ingresso]

1. Si scriva un programma che stimi l'entropia di ordine *m*, per valori di *m<=16 bit*.
2. Si valuti l'entropia di ordine *m* per vari file delle tre classi e si determini il valore massimo del rapporto di compressione (CR=bit originali/bit compressi) previsto dal primo teorema di Shannon

$$CR = m/H^{(m)}$$

3. Si tracci un grafico di CR al variare di *m* e lo si confronti con il valore sperimentale ottenuto comprimendo lo stesso file con un applicativo di codifica a dizionario.

4. Si commentino i risultati ottenuti.

Suggerimenti:

- si consideri il caso m=1 (il file viene inizialmente considerato come un flusso binario)
- si valuti l'entropia per valori di m in modo tale che sia semplice effettuare letture da file con granularità di sottomultipli o multipli di 8 bit (byte), ad esempio m=1,2,4,8,16
- Per stimare la frequenza statistica dei simboli occorre allocare un vettore di 2^m interi. Assicurarsi di non utilizzare valori troppo grandi di *m* rispetto alla memoria a disposizione.