Huffman kódok

Jelölések: Legyen A a diszkrét forrás K betűvel ($2 \leq K < \infty$). P_k jelöli az a_k betű valószínűségét, amelyekről feltesszük, hogy legfeljebb egy kivételével pozitívak.

Egy bináris fában két csúcs testvér (sibling) ha közös a szülőjük.

Huffman kód megkonstruálása Az algoritmus a következő: a kódot egy bináris fával fogjuk ábrázolni (Huffman kód prefix kód lesz), a levelek felelnek meg a kódszavaknak a szokásos módon. Három lépés van:

1) Az L a forrásábécé jegyeinek valószínűségeinek listája, ezek fognak megfelelni a fa leveleinek.

2) Vegyük a két legkisebb valószínűséget az L listából, a nekik megfelelő csúcsokat testvérekké tesszük: felveszünk egy köztes csúcsot, amely a szülőjük lesz. A szülő és a gyerekeik közötti éleket felcímkézzük 0-val illetve 1-gyel.

3) Helyettesítsük a két valószínűséget az összegeikkel, ez felel meg a köztes csúcsnak. Ha az új L egy elemű akkor megállunk, egyébként visszaugrunk a 2) lépésre.

Strukturális karakterizáció

Az így kapott bináris fa olyan, hogy minden csúcshoz hozzá van rendelve egy szám: a szülőhőz rendelt szám megegyezik a gyermekeihez rendelt számok összegével, ezzel ekvivalens, hogy a szám mgegyezik a megfelelő leveleken levő számok összegével. Ennél kicsit több is igaz.

Definició: Egy bináris fa *sibling* tulajdonságú ha minden csúcsnak (kivéve a gyökérnek) van testvére és a csúcsoknak létezik egy sorbarendezése, hogy a megfelelő valószínűségek monoton csökkenő sorrendben vannak és minden csúcs szomszédos a testvérével a sorrendben.

I. Megjegyzés: Az azonos nagyságú valószínűségek miatt a monoton csökkenő sorrend nem feltétlenül egyértelmű, ezért van csak létezés a definicióban.

II. Megjegyzés: A Huffman kódhoz tartozó fának triviálisan megvan a sibling tulajdonsága a konstrukció miatt, amelyben $2K-2$ elemű a lista.

I. Tétel: Egy bináris prefix tulajdonságú kód pontosan akkor Huffman-kód ha a kódfa sibling tulajdonságú.

Bizonyítás: Teljes indukcióval.

Megjegyzés: Egy l. szinten levő csúcs valószínűsége legfeljebb akkora, mint egy $l-1$. szinten levő csúcs valószínűsége.

Definició: Rendezett Huffman kód olyan Huffman kód, amelynél két testvér esetén a valószínűbb kapja a 0-s élt, a másik az 1-est.

Definició: Egy fa lexikografikusan rendezett kódfa, ha minden l-re az l. szinten levő
csúcsok valószínűségei legfeljebb akkorák, mint az $l-1$. szinten levő csúcsok valószínűsége, továbbá az l. szinten a valószínűségek monoton csökkennek a lexikografikus rendezésnek megfelelően.

II. Tétel: Egy bináris prefix kód pontosan akkor rendezett Huffman kód ha a kódfa lexikografikusan rendezett.

Bizonyítás: Teljes indukcióval.

Huffman kód redundanciája

$$r = \sum_{k=1}^{K} P_k n_k - H(P_1, \ldots, P_K)$$

ahol $H(P_1, \ldots, P_K) = -\sum_K P_k \log P_k$. Ismert, hogy optimális kódra $0 \leq r \leq 1$. Huffamn kódokra a redundancia felírható egy másik alakban is. A Huffman kód sibling tulajdonságú, azaz van egy monoton csökkenő valószínűségekből álló sorrend, hogy a $2k-1$. és a $2k$.-hoz tartozó csúcs szomszédosak. Legyen q_k a k. valószínűség a listán.

Ekkor a várható kódhosszúság következőképpen írható fel:

$$E(n) = \sum_{k=1}^{2K-2} q_k$$

Ez a következőképpen látható: minden q_i néhány P_j összege és P_k éppen n_k darab q_l összeadandója.

Hasonlóan átírható az entrópia is:

$$H(P_1, \ldots, P_K) = \sum_{k=1}^{K-1} (q_{2k-1} + q_{2k}) K(\frac{q_{2k}}{q_{2k-1} + q_{2k}})$$

ahol $K(x) = -x \log_2 x - (1-x) \log_2 (1-x)$. Ez valójában egy bináris fabeli "teleszkópos" összeg mivel

$$(q_{2k-1}+q_{2k}) K(\frac{q_{2k}}{q_{2k-1} + q_{2k}}) = -(q_{2k} \log_2 q_{2k} + q_{2k-1} \log_2 q_{2k-1} - (q_{2k-1}+q_{2k}) \log_2 (q_{2k-1}+q_{2k}))$$

Így ezt összegezve kapjuk, hogy

$$\sum_{k=1}^{K-1} (q_{2k-1} + q_{2k}) K(\frac{q_{2k}}{q_{2k-1} + q_{2k}}) = -\sum_k P_k \log_2 P_k + 1 \cdot \log_2 1 = H(P_1, \ldots, P_K)$$

Tehát

$$r = \sum_{k=1}^{K-1} (q_{2k-1} + q_{2k})(1 - K(\frac{q_{2k}}{q_{2k-1} + q_{2k}}))$$

Végül legyen $l \geq 1$ valamilyen szint, melyen a fa teljes (azaz $L = 2^l$ csúcs van a szinten), de van csúcs az $l+1$. szinten is. Ha $K > 2$ akkor ilyen l létezik. Legyen m

a legkisebb egész, amelyre a $2m-1$. csúcs az $l+1$. szinten van és legyenek $q'_1, \ldots, q'_L$ a valószínűségek az l. szinten. Ekkor az előző összeget szétvágva kapjuk, hogy

$$r = l - H(q'_1, \ldots, q'_L) + \sum_{k=m}^{K-1} (q_{2k-1} + q_{2k})(1 - K(\frac{q_{2k}}{q_{2k-1} + q_{2k}}))$$

Ha $K = 2$ akkor $l = 1$ mellett nincs második tag. Ezekkel az előkészületekkel könnyebb a következő tétel bizonyítása.

Tétel: Legyen P_1 a legvalószínűbb betű valószínűsége. Ekkor a Huffman kód redundanciájára teljesül, hogy

$$r \leq P_1 + \sigma$$

ahol $\sigma = 1 - \log_2 e + \log_2 \log_2 e \approx 0,086$. Ha $P_1 \geq \frac{1}{2}$ akkor

$$r \leq 2 - K(P_1) - P_1 \leq P_1$$

Bizonyítás: Ha $0 \leq x \leq \frac{1}{2}$ akkor $K(x) \leq 2x$. Ezt az egyenlőtlenséget alkalmazzuk a levágott részre:

$$\sum_{k=m}^{K-1} (q_{2k-1}+q_{2k})(1-K(\frac{q_{2k}}{q_{2k-1}+q_{2k}})) \leq \sum_{k=m}^{K-1} (q_{2k-1}+q_{2k})(1-\frac{q_{2k}}{q_{2k}+q_{2k-1}}) = \sum_{k=m}^{K-1} q_{2k-1}-q_{2k} \leq q_{2m-1}$$

Ez utóbbi egyenlőtlenség azért igaz, mert a q_i-k monoton csökkennek. Tehát

$$r \leq l - H(q'_1, \ldots, q'_L) + q_{2m-1}$$

Legyen n_1 a legrövidebb kódszó hosszúsága, ez ahhoz a betűhöz tartozik, amelynek P_1 a valószínűsége.

Először tegzük fel, hogy $P_1 \geq \frac{1}{2}$, ekkor $n_1 = 1$ és $l = 1$, azaz éppen azt kapjuk, hogy $r \leq 1-K(P_1)+q_{2m-1}$. Mivel $q_{2m-1} \leq 1-P_1$, így $r \leq 2-K(P_1)-P_1$. A második egyenlőtlenség $P_1 = \frac{1}{2}$ valamint $P_1 = 1$-re egyenlőséggel teljesül, így a konvexitás miatt ebben az intervallumban mindenhol teljesül. Ez $K = 2$-re is teljesül, így ezt az esetet elintéztük.

Továbbiakban legyen $n_1 > 1$. Ha minden kódszó ugyanolyan hosszú akkor legyen $l = n_1 - 1$, különben legyen $l = n_1$. Mindkét esetben $q_{2m-1} \leq P_1$. Legyenek az l. szinten a valószínűségek $q'_1, \ldots, q'_L$, melyekre $q'_1 \geq q'_2 \geq \ldots q'_L \geq q'_1/2$. Legyen Q azon

$q'_1, \ldots, q'_L$ választások halmaza, amelyek teljesítik a fenti egyenlőtlenséget, továbbá $\sum q'_i = 1$. Ekkor

$$r \leq l - \min_{\{q'_i\} \in Q} H(q'_1, \ldots, q'_L) + P_1$$

Mivel H konkáv függvény, így a minimum Q valamely extremális pontjában vétetik fel.
Q extremális pontjai azok a pontok, amelyekre létezik n, hogy $1 \leq n \leq L$, $q'_i = q'_1$ ha $i \leq n$ és $q'_j = q'_1/2$ ha $n < j \leq L$. Ekkor adott n-re $q'_1 = \frac{2}{L+n}$ azaz

$$\min_{\{q'_i\} \in Q} H(q'_1, \ldots, q'_L) = \min_{1 \leq n \leq L} \left[-\log_2 \frac{2}{L+n} + \frac{L-n}{L+n} \right]$$

Ha n-t nem csak az egészeken hagyjuk futni kapjuk, hogy a minimum $l - \sigma$ ahol $\sigma = 1 - \log_2 e + \log_2 \log_2 e$. Ezt behelyettesítve éppen a bizonyítandó állítást kapjuk.

Megjegyzés: A becslés egészen pontos. Ha $P_1 \geq \frac{1}{2}$ akkor a $(P_1, 1 - P_1, 0)$ valószínűségekkel ellátott forrás egyenlőséggel teljesíti a becslést. Ha a valószínűségek $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ akkor a redundancia 0.415, a becslés pedig 0.419-et ad.

A továbbiakban azt vizsgáljuk, hogy hogyan változik a redundancia ha protokol célokra megtartunk egy 1 vagy 2 hosszú kódszót.

Tétel: Minden véges forráshoz létezik egy prefix kód, amelynek redundanciája legfeljebb 1 és nem használ egy 2 hosszúságú kódszót.

Bizonyítás: Először megkonstruáljuk a Huffman kódot, majd vesszük a kevésbé valószínű csúcsot az első szinten és a második szintre pakoljuk, az új 2-hosszú kódszó lesz az ő testvére és felvesszük még a szülőjüket. Így ezen az ágon minden kódszó 1-gyel hosszabb lett. Legyen q_1 és q_2 az eredeti fa első szintjén a két valószínűség. Ekkor ha r volt az eredeti redundancia akkor az új r' redundanciára kapjuk, hogy $r' = r + q_2$. $l = 1$-re használva az előző bizonyítás második képletét kapjuk, hogy

$$r \leq 1 - K(q_2) + q_2$$

$$r' \leq 1 - K(q_2) + 2q_2 \leq 1$$

Utóbbi egyenlőtlenségben felhasználtuk, hogy $q_2 \leq \frac{1}{2}$. Ezzel bebizonyítottuk az állítást.

Megjegyzés: Az $(\frac{1}{2}, \frac{1}{2}, 0)$ valószínűségekkel ellátot forrásra a becslés egyenlőséggel teljesül. Megmutatható, hogy a fenti eljárás a nem használt 2-hosszú szó megválasztására optimális r' minimalizálása szempontjából.

Adaptív Huffman-kódolás

Ha nem ismerjük az egyes betűk valószínűségét akkor a forrás betűinek relatív gyakoriságát használjuk. A probléma az, hogy nincs mindig lehetőség arra, hogy végigvárjuk az egész betűsorozatot. Ezért a következőt csináljuk: minden N lépés után az addigi eloszlásra felépítjük a Huffman-fát és azt alkalmazzuk, ezt mind a kodoló, mind a dekodoló megcsinálja. Az elején valamilyen a priori eloszlást használunk (ha semmit sem tudunk a forrásról, akkor mondjuk az egyenletes eloszlást használjuk).

Nincs szükség a tényleges relatív gyakoriságokat kiszámolni, mivel ezeknek csak egymáshoz viszonyított arányuk fontos, ezért dolgozhatunk a gyakoriságokkal, azaz egész számokkal. Arra sincs szükség, hogy a Huffmann-fát mindenegyes alkalommal újra építsük, elég ha ágakat cserélünk (bár kis fáknál, azaz kevés betűnél ez édesmindegynek tünik).

Apró technikai trükk az öregbítés: mivel a betűk eloszlása változhat, ezért érdemes az új N betű előtt beszorozni a gyakoriságokat valamilyen $0 < \alpha < 1$-val, így az új betűk nagyobb súllyal számítanak. A megfelelő (N, α) pár kitalálása problémafüggő, attól függ, hogy az eloszlás várhatóan milyen gyorsan változik.