2.7. SISTEME DE COMPRESIE DE DATE CU CODARE PREDICTIVĂ

Sistemele cu predicție se folosesc în compresia de date începând din aceeaşi perioadă în care au început să fie folosite tehnicile de modulație în cod de impulsuri. Prima variantă a codării cu predicție a fost aşa numita Modulație Diferențială în Cod de Impulsuri (Differential Pulse Code Modulation - DPCM). In principiu, DPCM se aplică pentru semnale mesaj analogice, care urmează să fie transmise în linie ca o succesiune de valori binare obținute prin serializarea cuvintelor de cod obținute prin conversie analog numerică. Dacă conversia se face cu o precizie de m biți, iar frecvența de eşantionare este f_s , atunci numărul de biți transmişi într-un interval T este $M=Nm$, cu $N=Tf_s$. Numărul poate fi redus dacă în loc să transmitem valoarea fiecarui eşantion, transmitem doar valoarea diferenței față de eşantionul precedent. Pentru semnale lent variabile (ex.: semnal vocal) se poate presupune că diferența (cu semnul inclus) se poate reprezenta pe un număr k<m de biți; numărul total de biți ce va fi transmis în acelaşi interval T fiind $M' = m + (N-1)k$.

Schema de principiu pentru un astfel de codor se înscrie însa într-o categorie mai amplă de metode, numite metode de codare predictivă. In fig. 2.23 se prezintă schema de principiu a codorului:

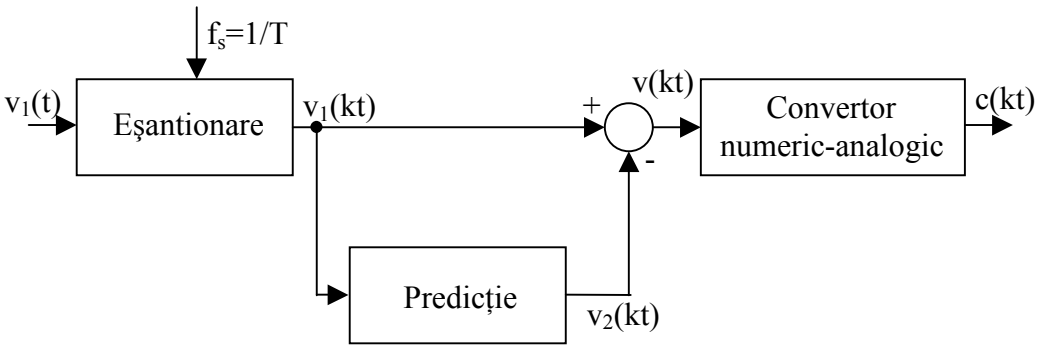

Fig. 2.23 Bloc de codare cu predicție

In schema de codare $v_2(kT)$ este valoarea curentă predictată a intrării $v_1(kT)$, calculată pe baza cunoaşterii valorilor precedente.

Decodarea, în sistemele de compresie cu predicție, se realizează în cazul general cu o schema analoagă celei de codare, prezentată în fig. 2.24.

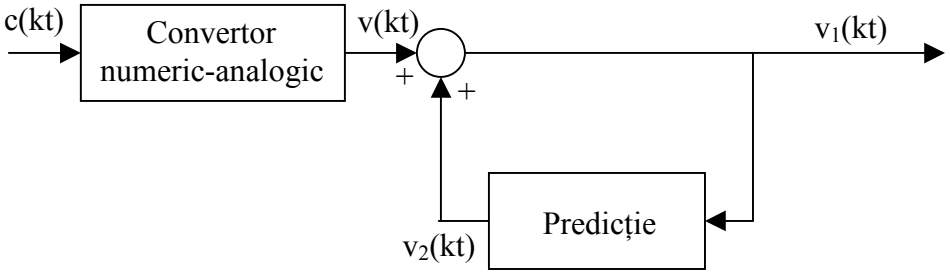

Fig.2.24. Bloc de decodare cu predicție

Deşi în schema din fig.2.23 conversia analog/numerică s-a efectuat după calcularea diferenței între valoarea curentă şi cea predictată, operația de conversie poate fi facută direct pentru mărimea de intrare, iar mărimea de predicție şi diferența sunt generate în formă discretă. In funcție de aplicație, predicția se poate rezuma la o singură cuantă de diferențiere față de eşantionul precedent (un pixel de imagine, un eşantion de semnal vocal), ceea ce conduce la metoda cunoscută sub denumirea Modulație Delta, care foloseşte pentru construirea valorii predictate creşterea sau scăderea cu o singură cuantă de nivel (delta). Modulația Delta nu este însă o metodă de compresie exactă (entropică), rata distorsiunii fiind cu atât mai mare cu cât rata de compresie (şi implicit valoarea cuantei delta) este mai mare.

Există însă o tehnică de compresie exactă pentru şiruri de caractere de tip predictiv foarte folosită, denumită *tehnica facsimil*. Este metoda utilizată la transmisiile prin fax şi constă în analizarea unei imagini linie cu linie. Se transmite integral prima linie (considerată un şir de caractere), după care se transmit din linia următoare doar caracterele care diferă, sub formă de grupe de două caractere, dintre care primul este un pointer (indicator) al poziției şi al doilea este caracterul cu care trebuie înlocuit cel vechi. Desigur, metoda devine eficientă doar dacă în fiecare linie cel puțin jumatate din caractere nu se modifică. Un exemplu tipic de aplicare cu succes a acestei metode este transmiterea imaginilor binarizate.

Codarea cu predicție se utilizează însă în compresie şi în variante foarte moderne, în scopul aplicării unui principiu de optimizare a reprezentării prin codare naturală propus de J. Rissanen cu denumirea Lungime Minimă de Descriere (Minimum Description Length – MDL) şi care se materializează prin construcția de coduri cu două părți[3],[37].

Înainte de a da o definiție a noțiunii de cod cu cea mai scurtă lungime a datelor, se va studia procesul de codare naturală pentru clasa de modele

$$M_k = \{P(x/\theta), \pi(\theta)\} \tag{2.16}$$

Pentru clasa $M_k = \{P(x/\theta), \pi(\theta)\}$ se vor considera parametri ce trebuie trunchiați cu o anumită precizie, ca fiind numere, de exemplu $\delta_j = 2^{q_i}$, unde q_i este numărul de zecimale luate pentru trunchiere. În conformitate cu teorema lui Shannon putem construi un cod prefixat *C* care asociază fiecărui vector de parametri astfel trunchiat o lungime de cod $L(\theta)$, dată de limita superioară integrală a lui $-\log P(x/\theta)$. Funcția *D* care decodifică pe *x* din cuvintele de cod concatenate

$$C(x,\theta) = C(\theta)C(x/\theta) \tag{2.17}$$

defineşte un sistem de codare cu un cuvânt de cod al lui *x* pentru fiecare valoare trunchiată a parametrului.

Pentru a obține codul cel mai scurt de date este normal să se caute valoarea parametrului care minimizează lungimea totală de cod

$$L(x,\theta) = C(\theta)C(x/\theta) \tag{2.18}$$

sau cu alte cuvinte

$$\min_{\theta} \left\{ -\log P(x/\theta) - \log \pi(\theta) - \sum_{j=1}^{k} \log \delta_j \right\} = L_{min} \tag{2.19}$$

.

Lungimea de cod L_{min} depinde de preciziile selectate pentru parametri care trebuie optimizați. Într-adevăr ultimul termen scade dacă se utililizează o precizie „mai brută“ (δ_j mai mare) în timp ce primul termen creşte în general, deoarece vectorul parametric trunchiat $\overline{\theta}$, poate devia mai mult de la valoarea optimală netrunchiată $\hat{\theta}$. Se presupune că dacă funcțile probabilistice modelate sunt netede astfel încât să poată fi dezvoltate în serii Taylor, se poate obține precizia optimă

$$L(x,\overline{\theta}) \le L(x,\hat{\theta}) + \frac{1}{2}\overline{\delta}'\Sigma\overline{\delta} - \sum_{j} \log \delta_j \tag{2.20}$$

unde prin Σ s-a notat matricea derivatelor de ordin 2 ale funcției $L(x,\theta)$ în raport cu parametrii evaluați într-un anumit punct de lângă $\hat{\theta}$ si prin $\overline{\delta}'$ a fost notat vectorul componentelor δ. Inegalitatea decurge din faptul că s-a evaluat forma pătratică din punctul în care diferența $\overline{\theta} - \hat{\theta}$ atinge maximul. În acest punct este convenabil a fi considerați logaritmii ca fiind logaritmi naturali. Atunci se minimizează termenul din dreapta

$$\sum \overline{\delta} = \overline{\delta}^{(-1)} \tag{2.21}$$

pentru preciziile optimale în cazul cel mai defavorabil, unde prin $\overline{\delta}^{(-1)}$ s-a notat vectorul de componente $1/\delta_j$. Această ecuație poate fi rezolvată numeric, prin tehnici numite de *cuantizare vectorială*. Dacă se notează preciziile optime cu $\hat{\delta}_j$, se obține membrul drept din

$$L(x,\overline{\theta}) \le L(x,\hat{\theta}) + \frac{1}{2}\overline{\delta}'\sum_{j}\overline{\delta} - \sum_{j}\log\hat{\delta}_j \tag{2.22}$$

minimizat

$$-\log\left(P(x/\hat{\theta})\pi(\hat{\theta})\right) + \frac{k}{2} - \sum_{j=1}^{k}\log\hat{\delta}_j \tag{2.23}$$

Dacă se presupune că $-\log P(x/\theta)$ creşte proporțional cu *n*, numărul observațiilor, elemente ale lui $S = \Sigma/n$ sunt de ordinul întâi indiferent de *n* şi atunci (2.21) implică $\hat{\delta}_j = c_j(n)/\sqrt{n}$, unde $a < c_j(n) < b$, iar expresia

$$-\log\left(P\left(x/\hat{\theta}\right)\pi\left(\hat{\theta}\right)\right)+\frac{k}{2}-\sum_{j=1}^{k}\log\hat{\delta}_j \tag{2.24}$$

se reduce la

$$MDL(k)=-\log\left(P\left(x/\hat{\theta}\right)\pi\left(\hat{\theta}\right)\right)+\frac{k}{2}\log n+O(k) \tag{2.25}$$

unde ultimul termen este de ordin k şi este dat de

$$O(k)=\frac{k}{2}+\sum_{j=1}^{k}c_j(n) \tag{2.26}$$

Pentru dimensiuni mai mari de eşantion, costul modelului este dominat de termenul al doilea.

Criteriul *MDL(k)*, introdus după cum am spus de Rissanen, permite evaluarea unei metode de compresie entropică în spiritul lucrărilor lui Shannon, care până acum nu puteau fi utililizate ca atare pentru a selecta clase stohastice de modele.

Folosind iniţial drept criteriu de selecţie *MDL(k)* poate fi utilizat şi pentru obţinerea celei mai scurte lungimi de cod pentru datele din clasa de modele M_k. Sistemul de codare

$$C\left(x/\theta\right)=C(\theta)C\left(x/\theta\right) \tag{2.27}$$

este redundant în sensul că fiecare secvenţă de date poate fi codată cu fiecare valoare parametrică, în timp ce într-un sistem de codare optimal ar trebui să folosim doar un singur cuvânt cod pentru fiecare secvenţă de date. Reducerea entropiei redundante se face ţinând seama de complexitatea stochastică a modelului[48].

Deoarece codul cu două părţi deja satisface condiţia de prefix, se poate substitui lungimea cuvântului de cod

$$\min_{\theta}\left\{-\log P\left(x/\theta\right)-\log\pi(\theta)-\sum_{j=1}^{k}\log\delta_j\right\} \tag{2.28}$$

prin inegalitatea lui Kraft

$$P'(x)=\sum_{C_i\in S_n(s)}2^{-|c_i|} \tag{2.29}$$

obţinând rezultatul

$$P'(x)=\sum_{\gamma}2^{-L(x,\gamma)}\prod_{j}\delta_j \tag{2.30}$$

unde suma este compusă din toate valorile $\hat{\theta}$ ale parametrilor trunchiaţi. Datorită distribuţiilor datelor de forma

$$P(x)=\int P\left(x/\theta\right)d\pi(\theta) \tag{2.31}$$

se obţine complexitatea stochastică a datelor corespunzătoare clasei de modele

$$M_k=\left\{P\left(x/\theta\right)\right\} \tag{2.32}$$

ca fiind

$$I(x/M_k) = -\log P(x) \tag{2.33}$$

În acest mod se poate considera $P(x/\theta) = f(x/\theta)$ o funcție densitate şi se obține complexitatea stochastică definită de

$$I(x/M_k) = -\log f(x) \tag{2.34}$$

Faptul că aceste lungimi de cod obținute prin mutarea unei redundanțe în sistemul de codare, definit de clasa modelelor, le împrumută acestora un sens natural de minimalitate.

Este uşor de remarcat că $I(x/M_k) = -\log P(x)$ este mai mic decât cel mai bun cod cu două părți. Suma

$$P'(x) = \sum_{\gamma} 2^{-L(x,\gamma)} \prod_{j} \delta_j \tag{2.35}$$

este clar mai mare decât oricare dintre termenii săi, inclusiv maximul.

Utilizând dezvoltarea în serie Taylor a probabilității maxime estimată $\hat{\theta}$, se obține

$$I(x/M_k) \approx -\log P(x/\hat{\theta}) + \frac{1}{2}\log\left|\hat{\Sigma}\right| \approx -\log P(x/\hat{\theta}) + \frac{k}{2}\log n \tag{2.36}$$

O asfel de comparație este în special convenabilă în legătură cu aşa–numitele familii conjugate de modele $f(x/\theta), \pi(\theta/\alpha)$ pentru care integrala $P(x) = \int P(x/\theta)\,d\pi(\theta)$ poate fi evaluată într-o formă adecvată.

Să revenim acum la schema de codare în două părți, în care codorul preia mai întâi cele mai potrivite valori ale parametrilor pentru a construi cu ele codul, apoi într-un preambul la şirul codificat, transmite decodorului informația. Dacă însă decodorul în loc să preia, calculează cele mai bune valori ale parametrilor din şirul anterior, utililizând un algoritm cunoscut de decodare, preambulul nu mai este necesar. Rezultatul este un proces de *codare predictivă* care furnizează o altă procedură de eliminare a redundanței codului diferită de formula sumă sau integrală din complexitatea stochastică[6],[34].

Fie mulțimea ordonată de date notată cu $x^n = x_1,\ldots,x_n$. Notăm cu $\hat{\theta}(x^t)$ probabilitatea maximă estimată a componentei *k* a vectorului parametric, obținută din şirul anterior x^t. Această estimare este trunchiată la o anumită precizie cunoscută de codor. Spre deosebire de codarea în două părți, nu e necesar să se optimizeze precizia, aceasta putând fi luată suficient de fină pentru ca parametrii să reprezinte corespunzător distribuțiile modelului. MDL estimează şi de asemenea minimizează lungimea anterioară a şirului $-\log P(x^t/\theta)$, şi deci condiția indusă

$$P(x_t | x^{t-1}, \hat{\theta}(x^t)) = P(x_t | \hat{\theta}(x^t)) \Big/ P(x^{t-1} | \hat{\theta}(x^t)) \tag{2.37}$$

este cea necesară pentru a coda x_t. Cu această strategie se obține lungimea codului

$-\log P\left(x_{t+1}\middle|x^t,\hat{\theta}\left(x^t\right)\right)$, care aplicată peste întreg *t,* dă lungimea totală a codului predictiv pectru secvența:

$$L\left(x\middle|M_k\right)=-\sum_{t=0}^{n-1}\log P\left(x_{t+1}\middle|x^t,\hat{\theta}\left(x^t\right)\right) \tag{2.38}$$

Schema descrisă anterior necesită anumite modificări din cauza unor deficiențe. Cea mai mare deficiență este posibilitatea ca probabilitatea simbolului $P\left(x_{t+1}\middle|x^t,\hat{\theta}\left(x^t\right)\right)$ să fie nulă sau egală cu 1. În distribuțiile discrete aceasta se se va întâmpla pentru fiecare următoare apariție a aceluiaşi simbol. De exemplu să presupunem că şirul anterior este 0 şi că simbolul următor este 1. Probabilitatea maximă calculată pentru şirul anterior estimează $\hat{\theta}(0)=1$, care atribuie simbolului următor, care este 1, probabilitatea 0. Această anomalie poate părea uşor de evitat. Trebuie doar restrâns domeniul în care este estimată probabilitatea maximă a următorului simbol în intervalul $(\varepsilon,1-\varepsilon)$[33].

Un alt defect al lungimii de cod predictiv este acela că datele trebuie ordonate. Există, desigur, o ordonare optimă pentru care lungimea de cod predictiv este minimizată, dar aflarea ei necesită examinarea a *n*! ordini diferite. Pentru şiruri lungi nu are prea mare importanță cum este ordonat şirul de date, dar pentru şiruri de lungime scurtă şi medie o ordonare bună a secvenței inițiale este foarte importantă.

Utilizarea principiului MDL în evaluarea algoritmilor de compresie

Faptul că principiul MDL crează inevitabil distribuții pentru parametri îl aseamănă cu tehnicile bayesiene, atunci când acestea sunt aplicabile. În forma sa cea mai pură filozofia Bayesiană este bazată pe principiul conform căruia datele sunt generate prin distribuție parametrică $f(x/\theta)$ şi conform căreia cunoştințele anterioare despre valoarea parametrilor pot fi reprezentate printr-o altă distribuție $\pi(\theta)$. Rezultatul $f(x/\theta)\pi(\theta)$, obținut prin normalizare integrală:

$$f(x)=\int f(x/\theta)\mathrm{d}\pi(\theta) \tag{2.39}$$

definește aşa numita distribuție posterioară $\pi(\theta)$, care poate fi interpretată ca o sinteză naturală a celor două surse de informație, datele anterioare şi cele experimentale. Cu toate că procesul, exprimat în teorema lui Bayes, pune în evidență ideea importantă a nesiguranței (despre valoarea parametrilor) condiționând (asupra datelor) sugestia că distribuția $\pi(\theta)$ captează cunoştințele anterioare într-o manieră adecvată, apar numeroase dificultăți de interpretare de câte ori parametrul este o constantă necunoscută. De aceea, un model care permite cea mai scurtă codare a datelor devine interesant şi pentru alte sarcini practice cum ar fi predicția. Se poate considera axiomă faptul că un model sau o clasă de modele, care permite ce mai scurtă codare posibilă a datelor, reflectă cel mai bine toate proprietățile datelor ce se doresc a fi estimate. Se poate demonstra că predictorii MDL pot fi în mod asimptotic optimali în probleme de regresie lineară sau procese ARMA. De fapt încă nu a apărut nici un caz în care în

procese, rezultatul obținut prin aplicarea MDL să fie anormal sau să aibă proprietăți nedorite

Apare totuşi o problemă care, conceptual, este serioasă, dar care poate fi depăşită în majoritatea aplicațiilor de interes practic, prin exercitarea predicției. Problema provine din dificultatea găsirii celor mai scurte lungimi de cod utilizate pentru a coda modelele şi clasele de modele. Modul de rezolvare al acestei probleme este părăsirea specificației limbajului informal de bază; în fiecare aplicație concretă se consideră numai o familie de clase de modele preselectate, fiecare dintre ele putând fi descrisă cu aproximativ aceeaşi lungime folosind cuvinte într-un anume limbaj şi notații matematice. Lungimea rezultată trebuie să fie mult mai scurtă decât numărul biților necesari în descrierea datelor, în relație cu clasa model, adică *complexitatea stochastică*. Modul în care modelele sunt colectate în clase este, bineînțeles, non-unic. Frecvent gruparea este făcută astfel încât toate modelele unei clase au acelaşi număr de parametri k, şi colecția acestui fel de clase m_k asupra unui număr de parametri defineşte o familie $M\{M_k / k \in K\}$.

Un exemplu este mulțimea tuturor polinoamelor, fiecare definind variația distribuției unei unități gaussiene. O clasă M_k atunci este definită de mulțimea tuturor polinoamelor de grad $k-1$. Cu condiția să se cunoască cum se calculează complexitatea stochastică în relație cu fiecare clasă, se poate calcula lungimea combinată optimală a codului.

$$L(x/M) = \min_k \{I(x/M_k) + L(M_k/M)\} \tag{2.40}$$

unde $L(M_k/M)$ este codul lungimii necesare pentru a specifica clasa M_k.

În cazuri simple se poate considera $L(M_k/M)$ = *log k*., dar în general, codarea claselor în cadrul familiei poate fi dificilă şi se cere o anumită doză de imaginație pentru a obține o bună estimare a $L(M_k/M)$. Faptul că sunt necesare astfel de precauții în legătură cu lungimea de cod pentru clasele de modele şi familii este ilustrat de exemplul extrem unde clasele de modele constau din unicul membru P, definit ca $P(x) = 1$, $P(z) = 0$, $z \neq x$, x însemnând secvența observată iar z o altă secvență de aceeaşi lungime. În mod clar complexitatea datelor observate x pusă în legătură cu această clasă este zero. Acelaşi tip de situație are loc în codarea predictivă efectuată pe baza estimărilor $\hat{\theta}(x')$ sau predicțiilor $\hat{x}_{t+1}$.

Principiul MDL sugerează o ierarhie de 3 niveluri a problemelor de modelare şi o dezvoltare a teoriei estimării de-a lungul următoarelor direcții:

- la cel mai de jos nivel se stabileşte familia de modele M_k şi numărul parametrilor, fixat, iar scopul este de a găsi valori parametrice bune sau optimale
- pe următorul nivel ierarhic se consideră o familie M cunoscută şi fixă şi în plus față de valorile parametrilor, va fi estimat şi numărul parametrilor
- pe al treilea nivel se doreşte să se găsească şi cea mai bună clasă de modele, în plus față de numărul de parametri şi valorile lor

Problema estimării complete sau a modelării, constând în ierarhia pe trei niveluri, este tratată de principiul MDL într-o manieră uniformă. Începând cu nivelul

cel mai înalt, se caută familia pentru a minimiza $L(x/M)$, si se spune că M^* este cea mai bună familie găsită.
În cadrul acesteia se găseşte cea mai bună clasă de modele pentru a minimiza

$$L(x/M) = \min_k \{I(x/M_k) + L(M_k/M)\} \tag{2.41}$$

care aici este specificată de numărul optim al parametrilor k. În sfârşit, se poate încerca să se găsească cel mai bun model în clasa $M^*_{k^*}$, definit de valoarea optimă a parametrului.