

一种对称/反对称正交平衡多小波的构造方法 及其在图像压缩中的应用研究^{*})

郑 武 余胜生 周敬利 陈加忠

(华中科技大学计算机学院 武汉430074)

摘 要 多小波同时具有正交、紧支、对称等对信号处理十分重要的特性,并获得很好的应用。目前大多数的多小波需要相应的预滤波技术。本文提出一种对称/反对称正交滤波器构造的方法,然后通过平衡的方法构造相应的平衡滤波器。最后,以图像压缩实验来测试此种滤波器的性能,实验结果表明,不需任何预滤波,SAOBM 的性能远远高于 DGHM 和 CL 多小波,接近或高于 Bi9/7。

关键词 图像压缩,多重小波,平衡,系数重排

An Approach to Construction of Symmetric-Antisymmetric Orthogonal Balanced Multiwavelet and its Research on Image Compression

ZHENG Wu YU Sheng-Sheng ZHOU Jing-Li CHEN Jia-Zhong

(School of Computer, Huazhong University of Science & Technology, Wuhan 430074)

Abstract Multiwavelet has orthogonality, compacted support and symmetry, these properties are very important for signal processing. However, most of Multiwavelets require related prefilter. An approach to construction of symmetric/antisymmetric orthogonal filter is proposed and its related balanced filter is constructed, no any prefilter is required. Experimental results prove its performance is superior to DGHM and CL multiwavelets, closer to or higher than Bi9/7.

Keywords Image compression, Multiwavelet, Balancing, Coefficients rearrangement

1 前言

单小波变换作为一种流行的多分辨方法,可同时进行时域和频域分析,具有时频局部化和多分辨特性,在信号处理方面获得广泛应用。然而,在许多应用场合,小波的对称性或反对称性对信号处理相当重要,但正交、紧支的单小波除 Haar 小波外都不具对称性^[1]。近年来,多小波的研究是一个活跃的方向。由于多小波能同时满足正交、紧支、对称等对信号处理十分重要的特性,因此受到众多相关人员的关注。Goodman 等人首先对多重小波进行了研究^[2],随后又给出了多重小波的一些特性^[3]。Donovan G 等用分形插值的方法首次构造出支撑为 $[0, 2]$ 的二重多小波 GHM^[4]。Chui 等构造出二重多小波 CL3、CL4^[5]。此外,还有不少学者构造出形式多样的二重多小波^[6~8]。总的来说,多小波在图像压缩、去噪、数字水印等信号处理领域中得到了很好的应用^[9~11]。

虽然 GHM 是二重对称小波,但生成它的正交矩阵滤波器并不对称,而在图像编码中,用作分解和重构的正交矩阵滤波器具有对称性才能很好地解决边界问题^[12]。另外生成 CL 系列二重对称小波的正交矩阵滤波器虽然对称,但频谱特性并不很好^[11]。本文提出了一种对称/反对称正交滤波器构造的方法。

对于单小波,高通滤波器可直接由低通滤波器得到。但对多小波而言,由于矩阵的乘法一般不具有交换性,因此高通滤波器不能直接由低通滤波器得到,需另外采用获得对称/反对称正交高通滤波器的方法。

将构造出来的对称/反对称正交滤波器组直接用于图像压缩,图像质量并不好^[12]。因此有学者提出相应的预滤波技术(Prefilter)^[13~16],以改善图像信号重构的质量。本文采用一种平衡技术(Balancing)来构造对称/反对称正交平衡滤波器

组。图像压缩时,不用预滤波也能获得满意的效果。

为了检验本文提出的对称/反对称正交平衡滤波器组在图像压缩中的性能,将其与文[16]进行比较,为了充分利用同级子带间的相关性,提高编码效率,本文提出了不同与文[8, 11]的数据重排方法。

实验结果表明本文提出的二重对称/反对称正交平衡滤波器组及数据重排的方法,能获得很好的图像压缩效果。

2 对称/反对称正交滤波器组的构造

$\phi_1, \phi_2, \dots, \phi_r (\phi_i \in L^2(R), i=1, 2, \dots, r)$ 是 r 元正交多小波,如果 $\phi_1(2^j x - k), \phi_2(2^j x - k), \dots, \phi_r(2^j x - k), j, k \in Z$ 形成一个 $L^2(R)$ 空间的一个正交基。多小波构造与 r 元多分辨分析相关联^[2]。 r 元多分辨分析 MRA 是 $L^2(R)$ 空间上一串闭子空间嵌套序列 $\{V_j\}$, 满足下列条件:

(a) $V_j \subset V_{j+1}, j \in Z$; (b) $\bigcap_{j \in Z} V_j = \{0\}, \bigcup_{j \in Z} V_j$ 在 $L^2(R)$ 中稠密, (c) $f \in V_j \Leftrightarrow f(2 \cdot) \in V_{j+1}$; (d) 存在 r 个函数 $\Phi = (\phi_1, \phi_2, \dots, \phi_r)$, 其整数平移系列 $\{\phi_j(\cdot - k): 1 \leq j \leq r, j \in Z\}$ 是 V_0 上的 Riesz 基。函数 $(\phi_1, \phi_2, \dots, \phi_r)$ 称为多尺度函数,由它们形成一个 $MRA(V_j)$ 。如果一组紧支的尺度函数,它们整数平移系列形成 V_0 的一个正交基,那么 V_j 就形成一个正交的 r 元 MRA。对于一个正交的 $MRA(V_j)$, 设 $V_{j+1} = W_j \oplus V_j, W_j$ 为 V_j 在 V_{j+1} 中的正交补。由条件(a)、(b)隐含 $W_j \perp W_{j+1} (j \neq k)$ 和 $\sum_{j \in Z} \oplus W_j = L^2(R)$ 。如果存在一组函数 $\Psi = (\psi_1, \psi_2, \dots, \psi_r)$, 其整数平移系列 $\{\psi_j(\cdot - k): 1 \leq j \leq r, j \in Z\}$ 构成 W_0 的一个正交基, $\Psi = (\psi_1, \psi_2, \dots, \psi_r)$ 是一组正交小波。

设 $\Phi = (\phi_1, \phi_2, \dots, \phi_r)^T$ 是一多尺度函数,则由条件(a)、(c)、(d)表明存在 $r \times r$ 的矩阵 H_k, G_k 满足细化方程(refinement equation):

^{*}) 本课题得到国家863高技术研究发展计划(2002AA11901010)资助。郑 武 博士研究生,研究方向为多小波与信号处理,图像与视频编码。余胜生 教授,博士生导师,研究方向为计算机网络与存储,多媒体数据处理与通信。周敬利 教授,博士生导师,研究方法为计算机网络与存储,多媒体数据处理与通信。陈加忠 博士研究生,研究方向为小波理论与信号处理,图像与视频编码。

$$\Phi(x) = \sum_{k \in \mathbb{Z}} H_k \Phi(2x - k) \quad (1)$$

相应的多重小波 $\Psi = (\psi_1, \psi_2, \dots, \psi_r)$ 的细化方程可表示为:

$$\Psi(x) = \sum_{k \in \mathbb{Z}} G_k \Psi(2x - k) \quad (2)$$

令 $H(\omega)$ 表示 $\{H_k\}$ 矩阵频率响应, 定义为:

$$H(\omega) = \frac{1}{2} \sum_{k \in \mathbb{Z}} H_k e^{-ik\omega} \quad (3)$$

相应的 $G(\omega)$ 表示 $\{G_k\}$ 矩阵频率响应, 定义为:

$$G(\omega) = \frac{1}{2} \sum_{k \in \mathbb{Z}} G_k e^{-ik\omega} \quad (4)$$

对(1)、(2)进行傅式变换得:

$$\hat{\Phi}(\omega) = H\left(\frac{\omega}{2}\right) \hat{\Phi}\left(\frac{\omega}{2}\right) \quad (5)$$

$$\hat{\Psi}(\omega) = G\left(\frac{\omega}{2}\right) \hat{\Psi}\left(\frac{\omega}{2}\right) \quad (6)$$

2.1 滤波器组正交的条件

如果

$$\int \Phi(x) \Phi^*(x) dx = \sigma_{k,0} I_r, k \in \mathbb{Z} \quad (7)$$

式(1)给出的多尺度函数是正交的^[2], 式(7)的频域表达式为:

$$\sum_{k \in \mathbb{Z}} \hat{\Phi}(\omega + 2k\pi) \hat{\Phi}^*(\omega + 2k\pi) = I_r.$$

分别生成 Φ, Ψ 的滤波器 H_k, G_k 的正交条件如下:

$$H(\omega) H^*(\omega) + H(\omega + \pi) H^*(\omega + \pi) = I_r \quad (8)$$

$$G(\omega) G^*(\omega) + G(\omega + \pi) G^*(\omega + \pi) = I_r \quad (9)$$

$$H(\omega) G^*(\omega) + H(\omega + \pi) G^*(\omega + \pi) = O_r \quad (10)$$

(*) 表示共轭转置, I_r, O_r 分别表示 $r \times r$ 单位矩阵和零矩阵。

正交条件(8)、(9)、(10)的时域表示式为:

$$\sum_{j \in \mathbb{Z}} H_j H_{j+2k}^T = 2\sigma_{k,0} I_r, k \in \mathbb{Z} \quad (11)$$

$$\sum_{j \in \mathbb{Z}} G_j G_{j+2k}^T = 2\sigma_{k,0} I_r, k \in \mathbb{Z} \quad (12)$$

$$\sum_{j \in \mathbb{Z}} H_j G_{j+2k}^T = O_r, k \in \mathbb{Z} \quad (13)$$

2.2 滤波器对称/反对称条件

设分别由细化方程(1)生成的尺度函数 $\Phi(x)$ (其支撑全为 $[0, N]$) 是对称/反对称的。不失一般性, 可设 $\Phi(x)$ 的第一个元素为对称, 第二个元素为反对称, 其他元素以此类推。即 $\Phi(x)$ 满足下式:

$$\Phi(x) = S_r \Phi(N - x)$$

其中 S_r 是 $r \times r$ 的对角阵, $S_r = \text{Diag}(1, -1, 1, -1, \dots, (-1)^{r-1})$ 。由上式可以推出: $H_k = S_r H_{N-k} S_r$ 。

对于偶数长度的低通滤波器, 则有:

$$H_k = S_r H_{2N-1-k} S_r \quad (14)$$

对于高通滤波器, 定义:

$$G_k = (-1)^{k+1} H_{2N-1-k} P \quad (15)$$

其中 $P = \text{antiDiag}(1, 1, \dots, 1)^T$,

如果 $H_k P H_{2N-1-k-2i}^T, k=0, 1, \dots, N-1-i, i=0, 1, \dots, N-1, i \in \mathbb{Z}$ 是对称矩阵, 则 H_k, G_k 满足正交条件(12)和(13)^[6]。

因此, 按照上述方法, 可以构造不同长度的对称/反对称滤波器组。

2.3 对称/反对称滤波器组的构造

下面以滤波器长度为3, 即 $N=2, r=2$ 为例来描述滤波器组的构造。设 H_k 有如下形式:

$$H_0 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, H_1 = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$$

$$\text{则由(14)可得 } H_2 = \begin{bmatrix} e & -f \\ -g & h \end{bmatrix}, H_3 = \begin{bmatrix} a & -b \\ -c & d \end{bmatrix}$$

由 $H(\omega)$ 的表达式, 分别令 $\omega=0, \pi$ 得:

$$H(0) = \begin{bmatrix} a+e & 0 \\ 0 & d+h \end{bmatrix}, H(\pi) = \begin{bmatrix} 0 & b-f \\ c-g & 0 \end{bmatrix}$$

根据(11)式有:

$$\begin{cases} a^2 + b^2 - a = 0 \\ c^2 + d^2 + g^2 + h^2 = 1 \\ cg - dh = 0 \\ b(d+h) + a(c-g) - g = 0 \end{cases} \quad (16)$$

由(16)式可得:

$$H_0 = \sin\theta \begin{bmatrix} \sin\theta & -\cos\theta \\ \sin\theta & \cos\theta \end{bmatrix}, H_1 = \cos\theta \begin{bmatrix} \cos\theta & \sin\theta \\ \cos\theta & \sin\theta \end{bmatrix}$$

因此, 由(14)、(15)式可得 $H_2, H_3, G_k (k=0, 1, 2, 3)$ 等滤波器的系数, 可以验证 H_k, G_k 满足正交条件。

3 滤波器组的平衡

多滤波器组可以说是一个多输入/多输出(MIMO)的系统。对于一维信号来说, 必须对信号进行向量化, 也就是说将原始信号变成2维信号。通过多相分解可以很容易地做到。

用多相表示 $H(\omega)$:

$$\begin{bmatrix} h_0(\omega) \\ h_1(\omega) \end{bmatrix} = H(e^{j\omega}) \begin{bmatrix} 1 \\ e^{-j\omega} \end{bmatrix}$$

如果低通道的 $h_0(\omega), h_1(\omega)$ 有不同的频谱特性, 即一个具有低通特性, 而另一个具有高通特性, 会引起通道的不平衡。在这种情况下, 多相分解的方法会造成低频信号与高频信号的混叠, 在对信号进行压缩处理的场合下, 重构的信号会产生强烈的振荡, 从而影响重构信号的质量。因此, 我们自然期望光滑信号完全由低频信号重构, 而通过高频滤波器的输出信号为0。在正交的情况下, 定义与低、高通分析相关的块 Toeplitz 矩阵:

$$L = \begin{bmatrix} \dots & H_0 & H_1 & H_2 & H_3 & \dots \\ & & H_0 & H_1 & \dots & \\ & & & \dots & \dots & \end{bmatrix}$$

$$H = \begin{bmatrix} \dots & G_0 & G_1 & G_2 & G_3 & \dots \\ & & G_0 & G_1 & \dots & \\ & & & \dots & \dots & \end{bmatrix}$$

那么就要求特征信号 $u = [\dots, 1, 1, 1, \dots]^T$ 满足:

$$L^T u = u, Hu = 0;$$

然而, 目前大多数已构造的多小波并不满足上述条件, 因此, 不少学者提出了预滤波技术使信号能够满足不平衡滤波器组的要求^[13~16]。

3.1 平衡条件及方法

从上述条件可以推出下列命题是等价的^[17]:

- a) $L^T u = u$; b) $[1, 1]$ 是特征值为1的 $H(0)$ 的左特征向量; c) $\Phi(0) = [1, 1]^T$ 。

因此要构造 SAOBM 滤波器, 关键是使 $u = [1, 1]$ 成为 $H(0)$ 的左特征向量。对于一正交矩阵 R , 使其满足:

$$[1, 1] R^T H(0) R = [1, 1];$$

也就是说, 要由 $H[k]$ 构造出成低通平衡滤波器, 必须找到一个正交矩阵 R 满足上式。

通过下式可构造相应的低通平衡滤波器:

$$H^b(\omega) = R^T H(\omega) R;$$

相应的尺度方程:

$$\Phi^b(\omega) = H^b\left(\frac{\omega}{2}\right) \Phi^b\left(\frac{\omega}{2}\right);$$

构造平衡滤波器因采用正交变换, $H^b(\omega)$ 将保持 $H(\omega)$ 的大部分性质。 $[1, 1]$ 是特征值为1的 $H(0)$ 的左特征向量; $\Phi^b(0) = [1, 1]^T$ 。 $H(\omega)$ 收敛, 则 $H^b(\omega)$ 也收敛。尺度函数的光滑性及

近似阶都未变化。此外平衡处理方式还保持了滤波器组的正交性,尽管失去了尺度函数的对称性,但通过构造相应的高通平衡滤波器:

$G^b(\omega) = G(\omega)R$; 小波函数的对称/反对称性能保持。证明如下:

$$\begin{aligned}\psi^b(\omega) &= G^b\left(\frac{\omega}{2}\right)\Phi^b\left(\frac{\omega}{2}\right) = G\left(\frac{\omega}{2}\right)R\left(\prod_{j=2}^{\infty}H(e^{j\omega/2^j})\right)\Phi^b(0) \\ &= G\left(\frac{\omega}{2}\right)R(R^T\left\{\prod_{j=2}^{\infty}H(e^{j\omega/2^j})\right\}R)\begin{bmatrix} 1 \\ 1 \end{bmatrix} = G\left(\frac{\omega}{2}\right) \\ &\quad \left\{\prod_{j=2}^{\infty}H(e^{j\omega/2^j})\right\}\Phi(0); (\Phi(0) = R\Phi^b(0)) = G\left(\frac{\omega}{2}\right) \\ &\quad \Phi\left(\frac{\omega}{2}\right) = \Psi(\omega)\end{aligned}$$

3.2 SAOBM 的构造

正交矩阵 R 由下式确定:

$$R = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

很显然 R 满足 $RR^T = I_r$ 。对称/反对称正交平衡滤波器组可由下列公式得到:

$$H_i^b = RH_iR^T$$

$$G_i^b = G_iR$$

由此完成对称/反对称正交平衡滤波器组的构造。图1、2、3、4分别为由上述方法构造出来的 SAOM 和 SAOBM 的频谱与尺度及小波函数。很明显,从频谱看,SAOM 的性能不及 SAOBM。

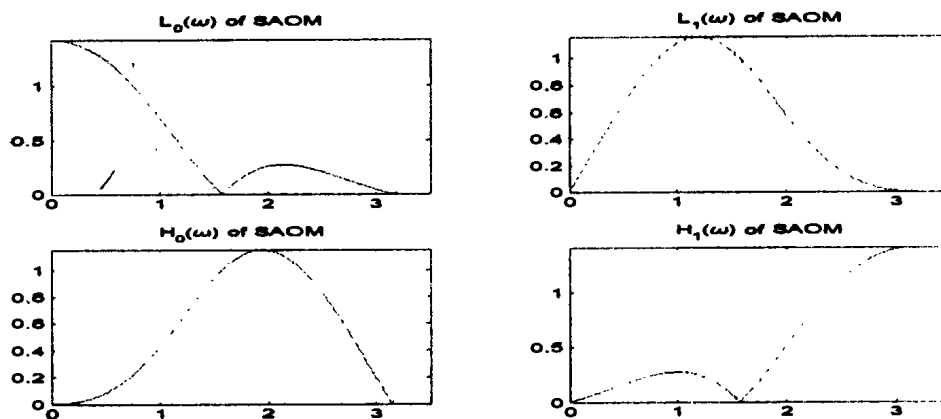


图1 平衡前 SAOM 滤波器的频谱

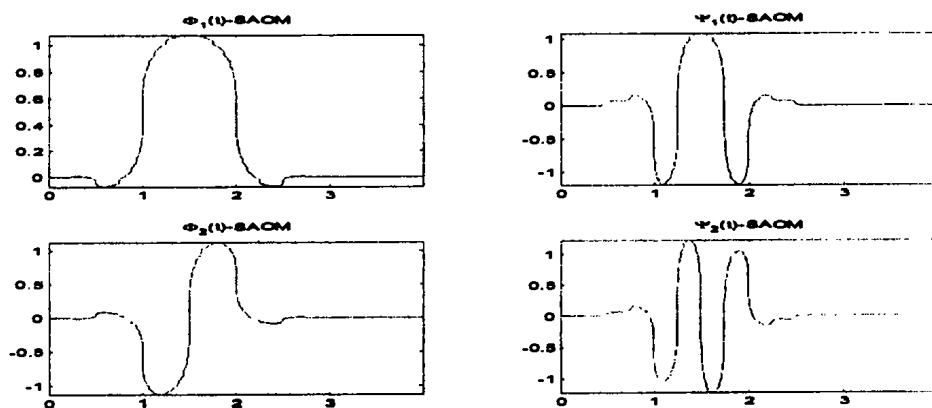


图2 平衡前 SAOM 滤波器的尺度及小波函数

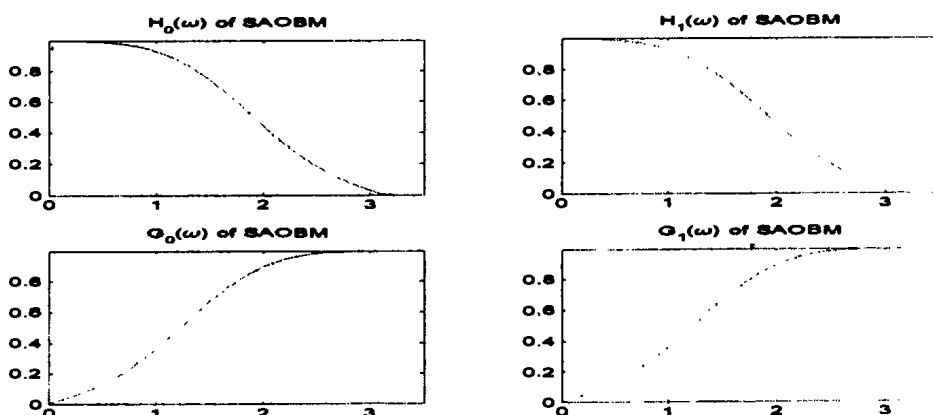


图3 SAOBM 滤波器的频谱

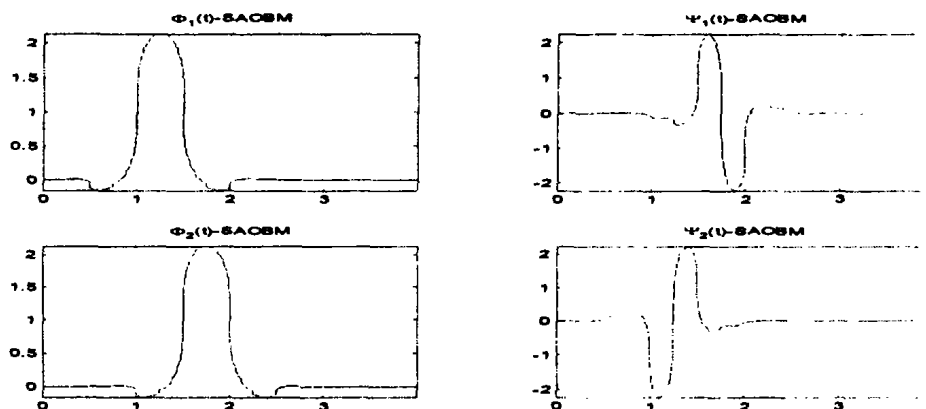


图4 SAOBM 滤波器的尺度及小波函数

4 基于 SAOBM 二重小波的图像编码研究

不同于单小波和一般的多小波分解,平衡多小波分解后的低频分量由4个部分组成:4个低频分量,如图5(a)所示,各子带系数之间的关系如图5(b)所示。由于SAOBM分解时,不需预滤波,在图像编码时,与单小波类似,可做4~5层的分解。

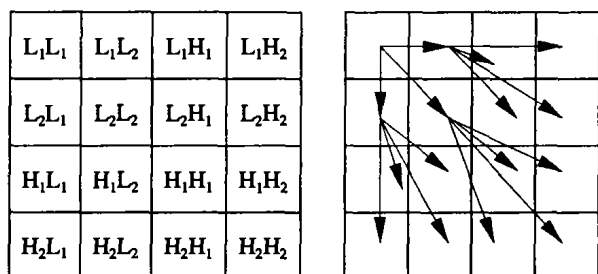


图5(a) 子带的分布结构

(b) 各子带系数的层次关系

为了充分利用 SPIHT 的性能,不同层同方向子带、同层各子带间必须要有很大的相关性。因此,为提高图像的编码质

量,有必要对 SAOBM 分解后的数据进行相应的处理。本文提出了一种平衡多小波系数重排方法(Balanced Multiwavelet Coefficients Rearrangement, BMCR),BMCR 不但能使同层各子带间更具相关性,而且不同层同方向子带的相关性并未受到影响,低频子带也进行同样处理(图6)。

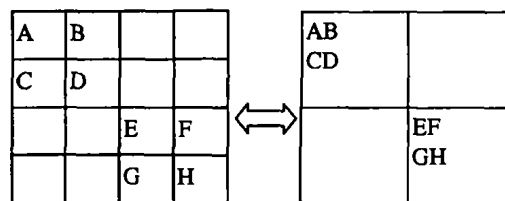


图6 对系数进行 BMCR

为了验证 SAOBM 性能,采用灰度图像是 512×512 的 Lenna 和 Barbara, SPIHT 编码,5层分解。实验结果分别为表1和表2。

表1 不同滤波器对图像 Lenna 编码结果比较(Binary Coding)

CPR	PSNR(dB)					
	CL Opt 1-3A	Db4	Balanced	DGHM Opt 1-3A	Bi9/7	SAOBM
8:1	38.65	38.74	38.05	39.24	39.82	39.71
16:1	35.02	35.19	34.31	35.76	36.72	36.41
32:1	31.68	31.85	30.95	32.22	33.60	33.26
64:1	28.86	29.02	28.20	29.15	30.35	30.28
128:1	26.42	26.51	25.91	26.53	27.34	27.56

表2 不同滤波器对图像 Barbara 编码结果比较(Arithmetic Coding)

CPR	PSNR(dB)					
	CL Opt 1-3A	Db4	Balanced	DGHM Opt 1-3A	Bi9/7	SAOBM
8:1	34.56	34.84	33.69	36.07	36.41	36.54
16:1	29.52	29.90	28.99	30.78	31.39	31.58
32:1	26.42	26.70	26.08	27.03	27.57	27.89
64:1	24.26	24.37	23.90	24.45	24.84	25.20
128:1	23.00	23.14	22.82	23.18	23.25	23.69

表1,表2中的 CL Opt 1-3A, Db4, Balanced, DGHM Opt 1-3A 的 PSNR 见文[16]。实验结果表明:

对于 Lenna, SAOBM 比 CL Opt 1-3A 高出约 $1 \sim 1.5$ dB, 比 DGHM Opt 1-3A 高出约 $0.47 \sim 1.1$ dB, 性能低于 Bi9/7 约 $0.07 \sim 0.3$ dB。

对于 Barbara, SAOBM 比 CL Opt 1-3A 高出约 $0.6 \sim 2$ dB, 比 DGHM Opt 1-3A 高出约 $0.5 \sim 0.9$ dB, 性能高于 Bi9/7 约 $0.1 \sim 0.44$ dB。

7约 $0.1 \sim 0.44$ dB。

结束语 多小波是近年小波研究的热点方向,在信号处理方面得到很好的应用。首先,本文提出了一种对称/反对称正交滤波器组的构造方法,接着构造相应的平衡滤波器组,最后对 SAOBM 进行验证。实验结果表明,SAOBM 滤波器具有良好的频谱特性,在图像编码中获得很好的效果,特别是对于复杂图像(如 Barbara),SAOBM 的性能优于单小波滤波器

Bi9/7.因此,多小波的理论及应用研究是有一定价值的。

参考文献

- Daubechies I. Ten lectures on wavelets. Philadelphia:SIAM Publ, 1992
- Goodman T N T, Lee S L. Wavelets in wandering subspaces. Trans. Amer. Math. Soc, 1993, 338(1):639~654
- Goodman T N T, Lee S L. Wavelets of multiplicity r . Trans. Amer. Math. Soc, 1994, 342(1):307~324
- Geronimo J S, Hardin D P, Massopust P R. Fractal functions and wavelet expansions based on several scaling functions. J. Approx. Theory, 1994
- Chui C K, Lian J. A study on orthonormal multiwavelets. Applied Numerical Mathematics, 1996, 20(3):273~298
- Shen L, Tan H H, Tham J Y. Symmetric-antisymmetric orthonormal multiwavelets and related scalar wavelets. Applied and Computational Harmonic Analysis, 2000(8):258~279
- Tan H H, Shen L, Tham J Y. New biorthogonal multiwavelets for image compression. Signal Processing, 1999(79):45~65
- Jiang Q T. On the design of multiresolution filter banks and orthonormal multiwavelet bases. IEEE Trans. Signal processing, 1998, 46(12):3292~3303
- Strela V, Heller P N, Strang G. The application of multiwavelet

- filterbanks to image processing. IEEE Trans. Signal Processing, 1999, 47(4):548~563
- Cotronei M, Lazzaro D. Image compression through embedded multiwavelet transform coding. IEEE Trans. Image Processing, 2000, 9(2):184~189
- Chen J Z, Zhou J L. A kind of symmetric orthogonal multiwavelets and the associated prefilter. Chinese Journal of computers, 2003, 26(6):701~707
- Strang G, Strela V. Short wavelet and matrix dilation equations. IEEE Trans. Signal Processing, 1995, 43(1):108~115
- Xia X G, Geronimo J S, Hardin D P, Suter B W. Design of prefilter for discrete multiwavelet transforms. IEEE Trans. Signal Processing, 1996, 44:25~35
- Xia X G. A new prefilter design for discrete multiwavelet transforms. IEEE Trans. Signal Processing, 1998, 46:1558~1570
- Hardin D P, Roach D W. Multiwavelets prefilter I: Orthogonal prefilter preserving approximation order $p \leq 2$. IEEE Trans. Circuits Syst. II, 1998, 45:1106~1112
- Attakitmongkol K, Hardin D P, Wilkes D M. Multiwavelet Prefilters-Part II: Optimal Orthogonal Prefilters. IEEE Trans. Image Processing, 2001, 10:1476~1487
- Lebrun J, Vetterli M. Balanced multiwavelets theory and design. IEEE Trans. Signal Processing, 1998, 46:1119~1125

(上接第138页)

在本系统中,为对文档进行训练,从 Yahoo!、网易、搜狐、中国期刊网等网站下载了分属于交通、教育、体育、军事4个类别共两百多篇已经明确了分类的文档作为测试。每个属性(即每个单字)都用一个4元组 $A = \{W, A, C, S\}$ 表示。其中, W 表示该字, A 是文档所属类别, C 是计数, S 是支持度。训练时,不是每个类别分别训练,而是把所有的训练文档全部输入,这样,就可以过滤掉那些在所有文档中出现频率都很高的字,然后,根据 TF-IDF 公式,计算出每个字的权重,并计算出支持度。

有了训练库,就可以从中提取出各分类的对应模式。模式可以根据权重或者支持度进行提取。权重或支持度的阈值既能由用户确定,也可由系统自动调整,也就是,如果系统分类错误的概率大于某值,则认为阈值太小,系统自动减小阈值。但是,阈值也不能太小,否则会导致干扰项太多。确定了阈值,就可以得到各个分类的模式,存放到模式库中。

系统对未知类型文档的处理,在特征字提取层是和训练文档一样处理的,不同的是,每个属性用三元组 $\{W, C, S\}$ 表示,中间结果存放在当前操作库中。与对训练库的处理一样,也要对当前操作库中的各个属性计算出权重和支持度。然后,分类器就可以对照模式库,输出该文档的可能分类。分类器采用 NB(Naive Bayes)分类方法。设 X 是待分类的文档,对于每个类别 C_i ,计算出 $P_i = P(X|C_i)P(C_i)$,令 P_i 最大的类别就是 X 的类别。对分类结果的准确性度量,可以用查准率(precision)^[9]和查全率(recall)^[9]来衡量。

$$precision = \frac{\sum correct_i}{\sum all_i}, (0 \leq i \leq j, \text{其中}, i, j \text{ 表示某个文档类别})$$

$$recall = \frac{|correct_j|}{|c_j|}, (0 \leq j \leq m, \text{其中}, m \text{ 是文档类别数})$$

其中, $correct_i$ 表示分类 i 中分类正确的文档数, $\sum correct_i$ 则是全部正确分类的文档数, all_i 是分类 i 的全部文档数, $\sum all_i$ 则是参加分类的全部文档数; $|correct_j|$ 是分类 j 中正确分类的文档数, $|c_j|$ 是属于分类 j 的全部文档数。

为了进行准确度测试,又从网上另外下载了43篇文档作为待分类文档输入到模型中,得到的结论是:分类结果的准确度与阈值关系密切,阈值太大或太小都会影响分类的精度。目前,最高的查准率和查全率都是在70%到80%之间,与一些使用 N -Gram (N 大于2)的分类方法^[4,5]相比(85%左右),略微低了一些。究其原因,一则,在中文文档中,大部分词条都是由

2个以上的词组成,如果采用了 N -Gram 方法,那些对文档区分度有很大权重的词条都可以被检索到,并且,通过设定阈值,可以过滤掉大部分干扰项,另一方面,这个模型还是有許多改进的余地。和 N -Gram 相比,它的速度更快,比较适用于需要即时分类的场合,如作为搜索引擎的搜索结果的快速分类器。

总结 本文提出了使用单字建立分类模型的一些初步构想和一个简单模型的实现。这种模型的优点在于:相对于词素解析,它不需要词典的支持;相对于 N -Gram,它的速度比较快。虽然准确度比其它分类方法略低,但70%以上的分类准确度还是令我们振奋的,下一步的工作,一是要研究阈值和准确度的关系,希望可以得到一个比较准确的阈值。另外,由于没有找到比较权威的非用词库,现有系统的非用词数据很不全面,需要进一步调整。不同类别的文档,它的非用词可能不尽相同,因此,对非用词进行分类,也是提高精度的手段。在上面的模型图中,虽然画出了反馈调整的模块,但目前还没完全实现,下一步需要完善它。另外,对于 Web 文档,是否可以利用它的超链接给我们提供的信息作为分类的参考?如何做?还有对于孤立点如何处理,这些都是我们下一步的研究目标。相信这种模型在经过进一步的改进后,精度还可以进一步得到提升。

参考文献

- Han Jiawei, Kamber M. Data Mining Concepts and Techniques. 2001
- Duda R O, Hart P E, Stork D G. Pattern Classification (Second Edition). 2003
- Wang K, Zhou S, Liew S C. Building hierarchical classifier using class proximity. In: Proc. 1999 Int. Conf. Very Large Data Bases (VLDB'99), Edinburgh, UK, Sept. 1999. 363~374
- 周水庚,等. 基于 N -Gram 信息的中文文档分类研究. 中文信息学报, 15(1):34~39
- 周水庚,等. 基于相邻字对信息的中文文档分类研究. 小型微型计算机系统, 2001, 22(4)
- 孙宏林,等. 浅层句法分析概述. <http://ling.ccnu.edu.cn/message/yyxlwx/collection-4/25qiancheng.htm>
- 常毅,等. 基于关键词表达式模型的文本自动分类系统的研究与实现. <http://www.software.ict.ac.cn/seminar/members/tjl/key-Expr>
- NII Test Collection for Information Retrieval Systems. <http://research.nii.ac.jp/~ntcadm/index-en.html>
- 王汉萍,等. 文本自动分类在搜索引擎上的应用. <http://www.ahcit.com/200305/27.doc>