

红外光谱小波压缩中最佳小波函数的选取

李梦龙 康 彬[#] 戚华溢 文志宁

(四川大学化学学院 成都 610064 [#]四川大学轻工与食品学院 成都 610065)

摘 要 研究了不同小波函数系对红外谱图压缩性能的影响。通过对 Aldrich 凝聚相傅立叶红外谱图库中 10607 张谱图的考察,把整个谱图库分为五类,分别选取一个最具代表意义的红外光谱,对 Daubechies 小波函数系、Coiflets 小波函数系、Symlets 小波函数系和双正交小波函数系在不同的小波函数阶数下进行实验。研究结果表明:一般文献中常用的 Daubechies 小波函数系并不具有最优压缩性能,双正交小波函数系才是最佳小波函数,其中最优化综合压缩性能的小波函数是 bior4.4 和 bior2.8。

关键词 压缩性能 红外光谱 小波函数类型 小波变换

The studies on the Choice of the Best Wavelet in Compressing IR by Wavelet Transformation

Li Menglong, Kang bin[#], Qi Huayi, Wen Zhining

(Department of Chemistry, Sichuan University, Chengdu 610064)

([#]College of Light Industry and Food Engineering, Sichuan University, Chengdu 610064)

Abstract This paper studied the influence of different type of wavelets on compressibility of infrared (IR) spectra. First, five representative IR spectra were selected from Aldrich library of FTIR spectra "Aldrich Condensed Phase". Then four different wavelet families such as Daubechies wavelets, Coiflets wavelets, Symlets wavelets and biorthogonal wavelets at different order are selected for investigation. The results indicated that Daubechies wavelets which are often employed in the literatures do not hold the best compressibility, but biorthogonal wavelets do it, of which bior 4.4 and bior2.8 hold the best general compressibility.

Key words Compressibility, IR spectra, Type of wavelet function, Wavelet transform

近年来,一种称为小波变换的数据处理技术被用于化学谱图数据的压缩中。小波变换不像傅立叶变换,函数基只能为三角函数,而是具有众多的不同的小波函数系可以选择,而且在同一小波函数系中还有不同阶数的小波函数。这一方面提高了解决不同问题的能力,另一方面使得对于不同性质的谱图数据,最优压缩性能的小波函数的选取成为一个必须考虑的问题。对压缩而言,应选择正交基^[1],如:Daubechies 小波函数系、Coiflets 小波函数系、Symlets 小波函数系和双正交的小波函数系等。虽然有文献讨论了最优压缩性能的小波函数,但大多仅仅是对 Daubechies 小波函数系。双正交小波函数系,也仅用于紫外可见光谱的压缩^[2]。章文军等^[1]对小波函数用于红外谱图压缩作了比较,但所选的小波函数仅限于 Daubechies 小波函数系和正交三

李梦龙 男,39岁,博士生导师,从事化学计量学方面研究。E-mail:liml@scu.edu.cn

国家自然科学基金资助项目(29877016)

2001-09-24 收稿,2002-04-01 修回

次 B-样条小波函数。Depczynski 等^[4]应用了被称为 Sturm-Liouville 小波, 避免了小波重建的边界效应, 但文中也仅与 Daubechies 小波函数系进行比较。其它的具有正交特性的小波函数系如: Coiflets 小波函数系、Symlets 小波函数系均未考虑。与其它化学谱图数据, 如: 紫外可见光谱或色谱数据相比, 红外光谱显得更为复杂, 特别是指纹区部分, 这使得红外光谱更不易压缩。而目前已知的化合物已有上千万种, 而且还在不断地合成或发现新的有机物和无机物, 标准红外谱图库也就越来越大。因此, 在保证红外谱图的主要特征基本不变的前提下, 如何对红外谱图进行有效地压缩, 较大地减少数据量, 对谱图的存储、检索及处理都是一项有意义的工作^[5]。其中对红外光谱数据压缩具有最优性能的小波函数的选取则是至关重要的一环。

1 小波压缩原理

从数学中的函数逼近论的观点来看, 压缩的本质是用尽可能少的小波基函数的加权求和项来最大限度地逼近原信号。如果基函数与原信号越相似, 则能用越少的求和项来逼近原信号, 在同样的恢复均方差下, 压缩比就越高, 压缩性能就越好。可见, 小波函数的选取对压缩是很重要的。小波变换不像傅立叶变换, 函数基只能为三角函数, 而是具有众多的不同的小波基函数。一般称函数 $\psi(x)$ 为小波母函数, 当且仅当其满足 $\int_{-\infty}^{+\infty} \psi(x) dx = 0$ 。这样的小波函数具有下列性质: 有足够强的光滑性, 即高次可微; 有局部性, 即函数本身与它的导数在无穷远处快速衰减为零, 甚至本身可以要求为紧支集; 有振荡性, 即具有充分多次的消失矩性。小波函数可以分为天然小波 (即有具体函数表达式的小波) 如: Morlet、Mexican; 无限正则小波如: Meyer; 正交紧支小波如: Daubechies 小波函数系、Symlets 小波函数系、Coiflets 小波函数系; 双正交小波如: 样条小波。但是只有具有正交特性的基函数才可能获得最大压缩比, 所以用于压缩的小波函数必须具有正交的特性。而 Daubechies 小波函数系、Symlets 小波函数系、Coiflets 小波函数系和双正交小波函数系均具有正交特性, 因此本文考察了这四类小波函数。

得到小波基函数权值的过程即被称为小波变换, 对于二进制小波, 实现小波变换的算法基本是采用 Mallat^[6]发明的 Mallat 塔式分解与重构算法, 浅显详细的说明可以参见 MATLAB Wavelet Toolbox 的使用手册^[7], 或写给化学工作者有关小波的一些 tutorials^[8]。在完成对原始信号的分解后, 虽然分解系数的个数还是等于或略大于原信号长度, 但这些系数的能量会很集中, 可以设定一阈值, 去掉大量绝对值小的系数, 只保留少数绝对值大的系数, 这样仍然可以在误差范围内重构出原始信号, 从而达到压缩数据的目的。

2 实验

实验中选取的小波函数类型如表 1 所示。

表 1 四种不同类型的小波函数系
Tab.1 Four different types of wavelet families

小波函数类型	小波函数阶数系
Daubechies	db1、db2、db4、db8、db10
Coiflets	coif1、coif2、coif3、coif4、coif5
Symlets	sym2、sym3、sym4、sym5、sym6、sym7
双正交小波	bior1.1、bior1.3、bior1.5、bior2.2、bior2.4、bior2.6、bior2.8、bior3.1、bior3.3、bior3.5、bior3.7、bior3.9、bior4.4、bior5.5、bior6.8

由于红外光谱本身的复杂程度等对压缩比有影响, 所以用来研究的红外谱图必须有代表

性, 这样找到的最优压缩性能的小波才能适用于整个红外谱图库的压缩。通过 Aldrich 凝聚相傅立叶红外谱图库中 10607 张谱图的考察, 把整个谱图库分为五类: (a) 简单的图谱; (b) 常规图谱; (c) 指纹区异常复杂的图谱; (d) 含有截断峰的图谱 (由于信号超量程所致); (e) 异常复杂图谱。分别选取一个最具代表意义的红外光谱: (a) triacontane, 99%; (b) 1-hexene, 99%; (c) anisole, 99%; (d) (+)-2,2,2-trifluoro-1-(9-anthryl)ethanol, 98+%; (e) phenyl phosphate, disodium salt dihydrate, 98%。五个谱图 (见图 1) 的具体参数如下: 波数范围为 $455.126 \sim 3995.852 \text{ cm}^{-1}$, 波数间隔为 7.714 cm^{-1} , 数据点共 460 点。Y 轴单位为吸光度 A 。

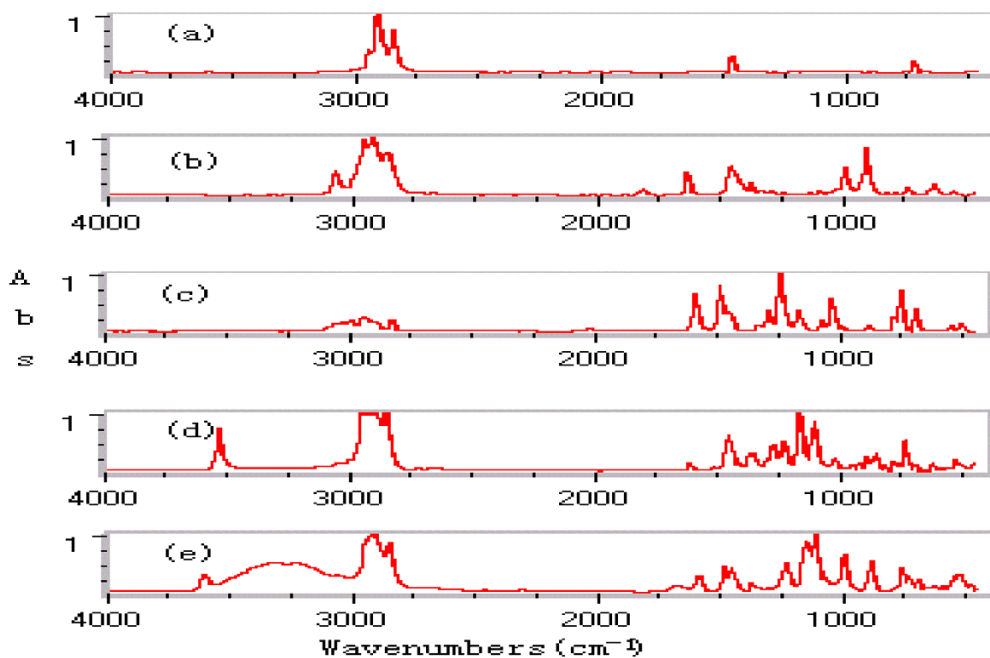


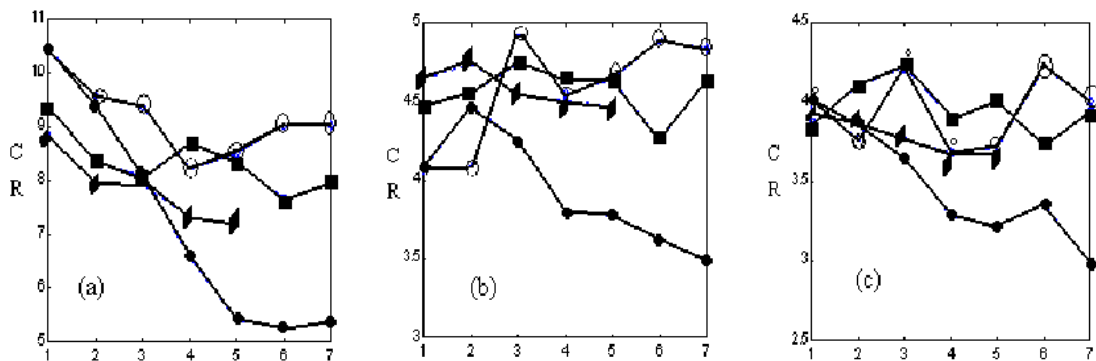
图 1 五个典型红外谱图

Fig.1 Five representative infrared spectra

根据笔者的前期工作, 采用预控均方差算法进行实验。均方差预控为 0.009。

3 结果与讨论

实验结果见图 2。



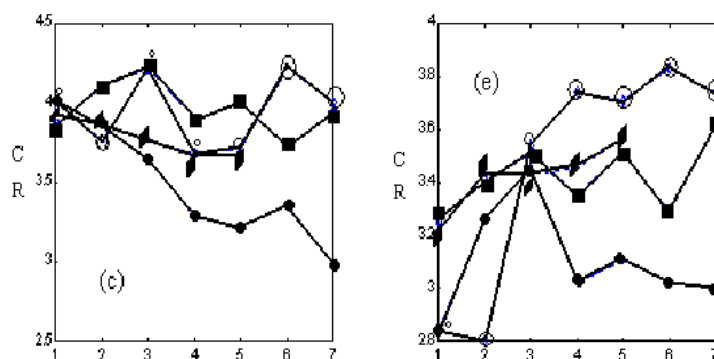


图 2 小波变换对不同谱图在不同小波函数系下得到的压缩比随小波阶数变化图

Fig.2 Compression ratios (CR) of different spectra vary with wavelet orders using different wavelet families
Daubechies 小波函数系(●), Symlets 小波函数系(■), Coiflets 小波函数系(◆), 双正交小波函数系(○)

图 2 中横坐标的刻度与实际小波函数阶数的关系如表 2。

从图 2 中可以看出: 在相同重建根均方差下, 不同类型的谱图数据, 不同的小波函数系, 不同的小波函数的阶数, 对压缩比均有影响。复杂程度不同的谱图得到的压缩比相差很大。谱图简单的如图 2(a), 最大压缩比可达 10 以上, 而谱图复杂的如图 2(d)、图 2(e)最大压缩比不到 4。同一小波函数系, 不同的小波阶数对压缩比的影响也很大不同, Daubechies 小波函数系受小波阶数影响最大。

表 2 横坐标的刻度值与实际小波函数阶数的关系

Tab.2 The relation between tick marks of X-axis and wavelets with different orders

Type of wavelet	Tick marks of X-axis						
	1	2	3	4	5	6	7
db	Harr	Db2	Db6	Db10	Db14	Db16	Db20
sym	Sym2	Sym3	Sym4	Sym5	Sym6	Sym7	Sym8
bior	bior1.1	bior1.5	bior2.8	bior3.3	bior3.7	bior4.4	bior6.8
coif	Coif1	Coif2	Coif3	Coif4	Coif5		

对简单谱图图 2(a), 无论什么小波函数系, 阶数越小, 压缩比越高, 最高为小波函数 Harr 和小波函数 bior1.1, 压缩比都为 10.45。前面已经讲到, 基函数与原信号越相似, 压缩比就越高, 正是由于小波函数 Harr 和小波函数 bior1.1 是最为简单的小波函数, 与简单的谱图图 2(a)相似, 才导致压缩比高, 随阶数的递增, 小波函数越复杂, 与谱图图 2(a)越不相似, 所以压缩比逐渐下降。对于 Daubechies 小波函数系, 选择小波函数 db16 作为压缩并不具有最佳压缩性能。

对常规图谱图 2(b), 不再是小波函数阶数最小的压缩比高, 对于不同的小波函数系具有最大压缩比的小波函数的阶数并不同。双正交小波函数系中 bior2.8, bior4.4, 表现最为出色, 有接近 5 的压缩比。Daubechies 小波函数系压缩性能仍然表现差劲。

对指纹区异常复杂的图谱图 2(c), 小波函数对压缩比的影响不大, 除 Daubechies 小波函数系外, 其余的小波函数系的压缩比都为 4 左右, 压缩比最高的还是双正交小波函数系中 bior2.8, bior4.4, 还有 sym4 表现也较为出色, Daubechies 小波函数系压缩性能仍不好。

对含有截断峰的图谱图 2(d), 情况与谱图图 2(c)相似, 小波函数对压缩比的影响不大, 除

Daubechies 小波函数系外, 其余的小波函数系的压缩比都为 3.2 左右, 双正交小波函数系和 Symlets 小波函数系的压缩性能相似, 都是最好的。Coiflets 小波函数系其次, 最差的还是 Daubechies 小波函数系。

对异常复杂图谱图 2(e), Daubechies 小波函数系和双正交小波函数系的压缩性能受小波阶数影响大, 最佳的小波函数的压缩比比最差的高 30% 左右, Coiflets 小波函数系和 Symlets 小波函数系的压缩性能对小波函数阶数不太敏感。双正交小波函数系中的 bior4.4, bior2.8 压缩性能最好。

总的来说, Coiflets 小波函数系和 Symlets 小波函数系的压缩性能对小波函数阶数不太敏感, 而 Daubechies 小波函数系和双正交小波函数系的压缩性能受小波阶数影响大。综合各类不同的谱图压缩情况, 双正交小波函数系表现最好, 其次是 Coiflets 小波函数系和 Symlets 小波函数, 最差的是 Daubechies 小波函数系。所有的小波函数中, bior4.4 和 bior2.8 对于所有的五个典型谱图综合压缩性能是最好的, 具体的数据如表 3 所示。

表 3 小波变换综合压缩性能最好的 bior4.4 和 bior2.8 压缩比结果

Tab.3 The results of compression ratios using bior4.4 and bior2.8 which hold the best compressibility

		(a)	(b)	(c)	(d)	(e)
bior2.8	offset	0.0555	0.0434	0.0360	0.0311	0.0333
	RRMSD	0.0091	0.0090	0.0090	0.0090	0.0089
	CR	9.39	4.95	4.22	3.13	3.54
bior4.4	offset	0.0592	0.0392	0.0369	0.0349	0.0378
	RRMSD	0.0090	0.0090	0.0090	0.0090	0.0090
	CR	9.02	4.89	4.22	3.17	3.83

要对上面的这一结论作严格的数学证明是非常困难的, 笔者仅对其作一浅显地分析。双正交小波函数系表现最好而 Daubechies 小波函数系表现最差, 其可能的原因在于两方面: 一是前面已经提到, 基函数的逼近能力越强, 压缩比就越高。User 等^[9,10]比较了双正交小波系和 Daubechies 小波函数系的逼近能力, 从数学上证明了对于光滑信号前者比后者有更好的逼近能力。二是小波函数系的对称性问题。Daubechies 小波函数系是不对称的, Symlets 小波函数系和 Coiflets 小波函数系也并非严格的对称, 但对称性都比 Daubechies 小波函数系好, 真正具有严格对称性的是双正交小波函数系。由于在信号处理中滤波器的对称性是一个非常重要的性质, 对称性好, 边界处理容易, 在重建时产生的相位失真小, 重建信号产生的畸变小。因此这也可能是双正交小波函数系优于 Daubechies 小波函数系的一个原因。

4 结论

用 Aldrich 凝聚相傅立叶红外谱图库中五个最具代表意义的红外光谱, 对四类小波函数系的压缩性能作了研究, 发现一般文献中常用的 Daubechies 小波函数系在红外谱图的压缩中表现一般, 并不具有最优压缩性能, 而双正交小波函数系表现良好, 其中 bior4.4 和 bior2.8 的压缩性能最好。这一结论对在保证红外谱图的主要特征基本不变的前提下, 对红外谱图进行有效地压缩, 较大地减少数据量有一定的意义。

参考文献

- [1] Kennneth K R. Digital Image Processing. Prentice Hall, Inc. 1996:343~345.
- [2] Ho H L, Cham W K, Chau F T et al. Comput. Chem., 1999, 23 (1):85~96.
- [3] 章文军, 许 禄, 刘胜雄. 高等学校化学学报. 1999,20 (4):534~538.
- [4] Depczynski U, Jetter K, Molt K et al. Chemom. Intell. Lab. Syst., 1999, 49(2):151~161.
- [5] Leung A K M, Chau F, Gao J B et al. Chemom. Intell. Lab. Syst., 1998, 43:69~88.
- [6] Mallat S G. IEEE Trans. Pattern. Anal., 1989, 11 (7):674~693.
- [7] Misiti M, Misiti Y, Oppenheim G et al. Wavelet Toolbox for Use with MATLAB, TheMathwork, 1996.
- [8] Alsberg B K, Woodward A M, Kell D B. Chemom. Intell. Lab. Syst., 1997, 37 (2):215~239.
- [9] User M. IEEE Trans. Singnal Processing, 1996, 44(3):519~527.
- [10] User M, Blu T. Proc. of SPIE, 1998,3458:14~21.