训练字典及其稀疏表示在近红外光谱法检测柴油中的应用

周扬^{*a,b} 戴曙光^a 吕进^b 刘铁兵^b 施秧^b

(^a 上海理工大学 光电信息与计算机工程学院 上海 200093)

(^b 浙江科技学院 近红外应用技术研究室 杭州 310023)

摘要 为提高柴油组分近红外法检测的精度,提出了一种基于训练字典稀疏表示下的建模方法并用于柴油十六烷值、沸点和芳烃总量的检测。该法先用柴油光谱结合 K 均值奇异值分解(K-SVD)算法完成对冗余字典的训练,再用正交匹配追踪算法(OMP)寻找柴油光谱在该训练字典下的稀疏表示系数,用该系数建立了柴油十六烷值、沸点和芳烃总量最小二乘预测模型。实验比对了训练字典、傅里叶字典和小波字典稀疏表示下的柴油组分预测模型性能,其中训练字典的表示系数建模性能最优且比其他两种字典的预测性能有较大幅度改进,验证了该法在近红外光谱检测建模领域推广能改善预测的准确度和稳健性。

关键词 近红外;柴油组分;稀疏表示;冗余字典;字典训练

Application of Training Dictionary and Sparse Representation in Diesel Property Detection Using Near-infrared Spectroscopy

Zhou, Yang^{*a,b} Dai, Shuguang^a Lv, Jin^b Liu, Tiebing^b Shi, Yang^b

(^a College of Optics & Electronics Engineering, University of Shanghai for Science and Technology, Shanghai 200093)

(^b Laboratory of NIR- Applications, Zhejiang University of Science and Technology, Hangzhou 310023)

Abstract It is an important method for predicting the diesel quality parameter by combination of near-infrared spectroscopy and chemometrics. In order to improve the accuracy of prediction model for diesel property, this paper puts forth a modeling method based on training dictionary and its sparse representation, and this method can test the cetane number, boiling point and total aromatics fast. The first stage of the method searches the particular redundant dictionary for diesel oil calibration samples, and the redundant dictionary can be trained by using K-means singular value decomposition algorithm (K-SVD). After finding of this redundant dictionary, it uses the orthogonal matching pursuit algorithm (OMP) to search the sparse representation coefficients of diesel spectroscopy under the training dictionary. This coefficients of the training dictionary, which is similar with the sparse representation coefficients of wavelet dictionary and fourier dictionary, is scale expansion of spectral characteristics that can effectively use to establish the pre-test partial least squares prediction (PLS) model of cetane number, boiling point and total aromatics for diesel. The spectra of diesel samples with different quality parameters from same factory are collected and the quality parameters of diesel samples are measured by ASTM certification method. The sparse representation coefficients of diesel spectroscopy under fourier dictionary and wavelet dictionary are figured in the experiments, inter-compared the performance of prediction model for diesel property under sparse representation coefficients of training dictionary, fourier dictionary and wavelet dictionary. Furthermore, the PLS model correlation coefficient (R) and standard deviation of prediction set (SEP) under sparse representation coefficients in training dictionary achieve 0.99 and 0.05 separately, it has optimal performance and also has considerable improvement to compare with the prediction performance of fourier dictionary and wavelet dictionary, which validated the efficiency of the method in predicting the diesel quality parameter using near-infrared spectroscopy and the proposed method can substantially improve the accuracy and stability of near-infrared spectroscopy prediction method.

Keywords near-infrared spectroscopy; diesel property; sparse representation; redundant dictionary; dictionary training

1 引言

近年来近红外光谱法逐步应用于柴油品质检测,国内外学者^[1~4]已成功利用该法测定了柴油芳烃含量、十六烷值、凝点、密度等各项指标,提高了分析效率、降

低了分析费用和改善了分析结果。由于柴油光谱结构复杂,图谱频带重叠严重,近红外光谱存在一定非线性,定量预测模型的回归效果受一定影响,对定量回归方法不断改进、运用波长特征选择和变换方法、将建模策略应用到建模过程^[5]是柴油光谱分析的主要研究方向。

* E-mail: zybuua@163.com

Received May 31, 2012; published August 6, 2012.

Project supported by the National Natural Science Foundation of China (No. 51075280), the Major Science and Technology Projects of Zhejiang Province (No. 2010C11060), the Natural Science Foundation of Zhejiang Province (No. Y4110235).

项目受国家自然科学基金(No. 51075280)、浙江省重大科技专项和优先主题计划(No. 2010C11060)和浙江省自然科学基金(No. Y4110235)资助。

为提高近红外光谱预测精度,文献^[6,7]成功将傅里叶变换和小波变换应用于近红外柴油分析过程,提高了模型的预测精度.这两种方法的实质是将光谱在傅里叶基或小波基上进行多尺度展开,展开式的部分分量含有化学特征信息.利用不同分量在变换域的稀疏特性实现光谱降噪、压缩和有效特征提取.随着稀疏表示理论的发展,傅里叶系数和小波系数作为字典已经不能更好地实现对信号的稀疏表示,新的冗余字典(过完备字典或过完备单位正交基)如 Curvelet, Contourlet, Wedgelet, Bandlet, Steerable wavelet^[8]等已经在信号稀疏表示领域广泛使用,其对原始信号的逼近能力要优于傅里叶系数和小波系数.但这些成果在近红外光谱分析领域尚未有相关报道,本文对柴油近红外光谱在特定冗余字典下的稀疏表示能否预测其物理特性和化学成分信息进行了初步研究.

对光谱信号的稀疏表示涉及冗余字典的选取,训练字典(即根据已知光谱信号训练得到的冗余字典)比上文提及的各种已知字典有着更好的稀疏表示能力. MOD, MAD, FOCUSS, K-SVD 等^[9]字典训练算法已具备利用近红外光谱完成冗余字典的训练能力,在此基础上,本文提出了一套使用柴油近红外光谱在训练字典稀疏表示下的回归校正方法,该法预测了柴油的十六烷值、沸点和芳烃总量.对比傅里叶和小波分解下的建模方法,研究了柴油近红外光谱在冗余字典的稀疏分解能否显著提高柴油物理化学特征预测性能.

2 方法

2.1 近红外光谱的稀疏表示

在光谱信号的有限维空间 R^n 中,近红外光谱 $y \in R^n$ 可由一组单位正交稀疏基 $D=[d_1 d_2 \cdots d_k]$ 线性表示, $y=Dx$, $d_k \in R^n$. 因其稀疏性,表示系数 $x \in R^k$ 中只有少数几个非零元素. 矩阵 D 中列向量代表稀疏基,由于 $k > n$,则认为 D 是冗余的(过完备、超完备)的,称为冗余字典. 在已知 y 和 D 的情况下,直接求解 x 是病态的. 文献^[10]将上述问题转换为式(1)和式(2)所示的优化问题.

$$\bar{x} = \arg \min_x \|x\|_0 \quad \text{s.t.} \quad \|y - Dx\|_2 < \varepsilon \quad (1)$$

$$\min_x (\|y - Dx\|_2) \quad \text{s.t.} \quad \|x\|_0 < T_0 \quad (2)$$

其中 $\|x\|_0$ 表示 x 的 l_0 范数,即 x 中非零元素的个数. 可用正交匹配追踪算法(OMP)^[11]对此优化问题求解. 该算法首先初始化残差 $r_0=y$ 、清空索引集 $A_0=\emptyset$ 和重组字典 $D_0=\emptyset$. 接下来重复步骤一和步骤二直到满足迭代停止条件.

步骤一:找出残差 r_i 与字典中的某个原子 d_j (字典中的列称字典的原子)内积最大值所对应的标号

$\lambda_i = \arg \max_{j=1,2,\dots,k} |(r_i, d_j)|$, 并更新索引集 $A_i = A_{i-1} \cup \{\lambda_i\}$ 和重组字典 $D_i = [D_{i-1}, d_{\lambda_i}]$.

步骤二:由最小二乘法得到当前光谱的最优逼近 $x'_i = \arg \min_x \|y - D_i x\|_2$ 并更新残差 $r_i = y - D_i x'_i$.

2.2 冗余字典稀疏表示回归算法

在完成对近红外光谱冗余字典稀疏表示系数 x 的求解后,将 x 系数代替光谱 y 与该光谱对应样本的化学、物理参数建立回归预测模型是提高近红外光谱法预测精度的有效方法,傅里叶变换和小波变换的就是利用傅里叶和小波两种字典,对光谱进行分解,提高了光谱的分辨率,照此思路,定义 Y^{cal} 为样本校正集近红外光谱矩阵, Y^{tst} 为验证集光谱矩阵, C^{cal} 为校正集化学物理参数测量值, C^{pred} 为验证集样本模型预测值,光谱冗余字典稀疏表示系数用于建模预测样本化学物理参数方法如图1所示.

图1 光谱冗余字典稀疏表示下回归预测

Figure 1 Model of sparse dictionary and spectrum representation

用冗余字典 D 对校正集光谱 Y^{cal} 和验证集光谱 Y^{tst} 进行稀疏分解,求得其校正集稀疏表示系数 X^{cal} 和验证集稀疏表示系数 X^{tst} ,利用 X^{cal} 和 C^{cal} 建立待测参数的偏最小二乘(PLS)^[12]预测模型,将 X^{tst} 作为模型输入,经计算得 C^{pred} 为光谱法对验证集样本某一物理化学参数的预测值.最后可用验证集光谱物理化学参数的预测值与标准方法的测量值比较来评价光谱法的准确性.

2.3 字典训练算法

文献^[13]已经证明小波和傅里叶两种字典对信号稀疏表示的性能并不是最优的,目前也尚无文献报道现有的某种确定性冗余字典适合近红外光谱稀疏分解,寻找能对光谱信号起增强作用的冗余字典 D 对能否提高模型的预测精度是关键. 文献^[14]提出一系列字典训练算法,即可利用校正集光谱 Y^{cal} 和验证集光谱 Y^{tst} 训练出适应其稀疏分解的冗余字典 $D=[d_1 d_2 \cdots d_k]$,本文采用的是性能较优 K 均值奇异值分解(K-SVD)算法.

设 Y 为 Y^{cal} 和 Y^{tst} 的光谱组合 $Y=[Y^{\text{cal}}, Y^{\text{tst}}]$, 其每一列代表一个光谱的样本,共计有 N 个样本. K-SVD 算法的目标就是在已知 Y 的前提下使目标函数最小:

$$\min_{D, X} \{\|Y - DX\|_2^2\} \text{ s.t. } \forall i, \|x_i\|_0 < T_0 \quad i=1, 2, 3, \dots, N \quad (3)$$

K-SVD 算法的每一次训练过程含稀疏分解(sparse coding)和字典更新(codebook update)两个步骤. 算法过程如图 2 所示.

图 2 K-SVD 算法

Figure 2 The K-SVD algorithm

稀疏分解过程用 OMP 算法将每个样本光谱按式(4)逐一求解其稀疏表示系数.

$$\min_{x_i} \{\|y_i - Dx_i\|_2^2\} \text{ s.t. } \forall i, \|x_i\|_0 < T_0 \quad i=1, 2, 3, \dots, N \quad (4)$$

字典更新过程每次只更新字典中的一个原子, 其他原子及对应稀疏表示系数固定. 此时目标函数的惩罚项由式(4)变换为式(5)所示形式, 其中 $E_k = (Y - \sum_{j \neq k} d_j x_j^j)$ 表示去掉第 k 个原子 d_k 后系数表示的误差.

$$\begin{aligned} \|Y - DX\|_2^2 &= \left\| Y - \sum_{j=1}^K d_j x_j^j \right\|_2^2 \\ &= \left\| (Y - \sum_{j \neq k} d_j x_j^j) - d_k x_k^k \right\|_2^2 = \|E_k - d_k x_k^k\|_2^2 \end{aligned} \quad (5)$$

定义 $\omega_k = \{i | 1 \leq i \leq N, x_T^k(i) \neq 0\}$ 为所有光谱样本 $\{y_i\}$ 在稀疏分解时用到原子 d_k 的索引群. 定义 Ω_k 为 $N \times |\omega_k|$ 矩阵, 在 $(\omega_k(i), i)$ 处值为 1, 其它都为 0. 定义 $x_R^k = x_T^k \Omega_k$, $Y_R^k = Y_T^k \Omega_k$, $E_R^k = E_T^k \Omega_k$, 三者分别为 x_T^k , Y_T^k , E_T^k 去零收缩后结果, 式(5)可进一步转化为:

$$\|E_k \Omega_k - d_k x_T^k \Omega_k\|_2^2 = \|E_R^k - d_k x_R^k\|_2^2 \quad (6)$$

将 E_R^k 进行奇异分解(SVD), 则 $E_R^k = U \Delta V^T$, 取 U 的第一列作为新的原子 d_k' , 取 V 的第一列和 $\Delta(1,1)$ 乘积作为新的系数矢量 $x_R^{k'}$. 不断重复稀疏分解和字典更新两个步骤, 直到迭代次数或稀疏表示误差达到停止条件.

3 结果与讨论

3.1 光谱稀疏度对模型预测结果的影响

在冗余字典训练过程中, 需选取合理迭代停止条件使模型的预测能力达到最优, 通常利用式(3)中的稀疏度 T_0 作为停止条件, T_0 表示光谱稀疏表示系数中非零分量的个数. 在柴油近红外光谱中不可避免地混入了噪声, 稀疏降噪理论^[9]证明了在光谱的稀疏表示中并不是字典中原子的线性组合越接近原始光谱信号越好, 由于光谱中噪声的不确定性, 实验设置 T_0 参数在 6 至 15 范围内对十六烷值、沸点和芳烃总量模型预测结果进行观察. 图 3 给出了在不同稀疏度下十六烷值、沸点和芳烃总量模型的校正集“留一”交叉验证标准差(SECV)和预测集标准差(SEP).

分析图 3 可以发现, 当稀疏度为 11 时, 十六烷值、沸点和芳烃总量模型的 SECV 和 SEP 具有相同的局部极小点. 该现象表明在模型应用时, 由于预测结果(SEP)的未知性, 对校正集采取交叉验证, 可利用交叉验证标准差(SECV)来确定模型最佳预测结果下的稀疏度 T_0 .

3.2 训练字典与小波、傅里叶字典表示性能比较

实验同时用 sym6 小波基(5 层分解)和傅里叶基(FT)对柴油光谱进行分解并用各表示系数分别建立了柴油十六烷值、沸点和芳烃总量的偏最小二乘预测模型, 与训练字典稀疏表示稀疏的偏最小二乘模型预测性能对比. 训练字典的初始采用离散余弦字典, 由于受训练样本数目要大于原子数的限制, 初始原子数为 100. 如表 1 所示, 十六烷值、沸点和芳烃总量三种基于训练字典稀疏表示下的预测模型相关系数($R=0.94, 0.99$ 和 0.99)最大和预测集标准差($SEP=0.29, 1.29$ 和 0.05)最小, 较小波、傅里叶字典分解下的模型, 预测性能有较大幅度的提升, 验证了训练字典对柴油近红外光谱具有较好的表示能力和稀疏分解能对光谱信号的化学物理特征有着更优的展示能力, 大幅度提高了柴油组分近红外法预测的准确度.

4 结论

实验验证了光谱训练字典稀疏表示法对柴油十六烷值、沸点和芳烃总量的建模预测性能远远优于小波和傅里叶变换建模方法, 提高了近红外光谱快速检测柴油品质的精度, 对油品在线快速检测具有参考价值. 但目前对光谱稀疏表示的性能评价除用校正集预测标准差外无其他有效评价手段, 该过程涉及多次训练和建模寻

图3 不同稀疏度下的模型预测结果

Figure 3 Prediction results under different sparsity

表1 不同分解表示下的柴油组分预测结果

Table 1 Prediction results of diesel under different decomposition

Ingredient method	Cetane number		Boiling point		Total aromatics	
	R	SEP	R	SEP	R	SEP
FT	0.61	3.71	0.88	13.03	0.96	0.42
WT sym6	0.60	4.14	0.89	11.11	0.92	0.95
Sparse	0.94	0.29	0.99	1.29	0.99	0.05

优过程,增加了计算量和复杂度.在字典训练过程中,样本数量要大于冗余字典原子数,导致字典训练时所需校正集样本过多,这是该法推广应用的一个限制.

在信号稀疏表示领域,匹配追踪算法和冗余字典训练算法改进是目前研究热点,大量文献^[15~17]提出性能更优的改良方法,提高了信号稀疏分解的分辨率.由于稀疏表示尚未在柴油光谱分析领域应用,本文只采用了经典的OMP和K-SVD算法作了初步尝试.随着研究的深入,新的研究成果将对光谱在训练字典下的稀疏表示性能进一步优化,不断提高近红外光谱模型预测的准确度和稳健性.

5 实验部分

取同一厂家、同一生产时间不同十六烷值、沸点和芳烃总量的柴油样本540个,其中510个作为校正集,30个作为验证集.校正集光谱采用Nicolet(尼高力)6700型傅里叶变换红外光谱仪扫描,光谱波长范围750~1550 nm(对应波数13333~6452 cm^{-1}),分辨率为0.4 cm^{-1} ,数据间隔6 nm,每个样品扫描10次.图4为校正集柴油

样本的近红外傅里叶变换光谱,测量结果与文献[2, 3]报道一致.柴油含氢基团(如C—H, O—H, N—H等)振动倍频、组合频的光谱位于近红外区,吸收波长与吸收强度是不同基团结构和组成信息的反映,适合碳氢有机物质的组成与性质测量.

柴油十六烷值理化值采用ASTM D613规定方法测定,范围为40.7~60.沸点理化值采用ASTM D86规定方法测定,范围为197~293 $^{\circ}\text{C}$.芳烃总量理化值采用ASTM D518规定方法测定,含量范围为14.3%~44.2%.实验用算法程序采用Matlab 2010a软件编写,偏最小二乘法调用美国Eigenvector Research公司的PLS工具箱内置函数.

图4 柴油样本的近红外光谱

Figure 4 NIR spectra of diesel samples

References

- [1] Ozdemir, D. *Pet. Sci. Technol.* **2008**, 26, 101.
- [2] Xu, G.-T.; Liu, Z.-L.; Yang, Y.-R.; Shen, S.-K.; Lu, W.-Z. *Acta Pet. Sinica* **2002**, 18, 65. (徐广通, 刘泽龙, 杨玉蕊, 沈师孔, 陆婉珍, 石油学报(石油加工), **2002**, 18, 65.)
- [3] Yuan, H.-F.; Chu, X.-L.; Lu, W.-Z.; Chen, W.-M.; Guan, X.-H.; Cao, W.; Zhang, C. *Chin. J. Anal. Chem.* **2004**, 32, 255. (袁洪福, 褚小立, 陆婉珍, 陈卫民, 关新虎, 曹巍, 张传, 分析化学, **2004**, 32, 255.)
- [4] Fabiano, B. G.; Celio, P. *Anal. Chim. Acta* **2010**, 670, 92.
- [5] Chu, X.-L.; Yuan, H.-F.; Lu, W.-Z. *Chin. J. Anal. Chem.* **2008**, 36, 702. (褚小立, 许育鹏, 陆婉珍, 分析化学, **2008**, 36, 702.)
- [6] Chu, X.-L.; Yuan, H.-F.; Lu, W.-Z. *Prog. Chem.* **2004**, 16, 528. (褚小立, 袁洪福, 陆婉珍, 化学进展, **2004**, 16, 528.)
- [7] Luiz, A. P.; Roberto, K. H. G.; Mário, C. U. A. *J. Braz. Chem. Soc.* **2011**, 22, 179.
- [8] Rubinstein, R.; Bruckstein, A. M.; Elad, M. *Proc. IEEE* **2011**, 98, 1045.
- [9] Elad, M.; Aharon, M. *IEEE Trans. Image Proc.* **2006**, 15, 3736.
- [10] Allan, P. J. *Approx. Theory* **2011**, 163, 388.
- [11] Zakria, H.; John, S. T.; David, R. H.; Charanpal, D. *IEEE Trans. Inf. Theory* **2011**, 57, 5326.
- [12] Julio, C. L. A.; Claudete, B. H.; Ronei, J. P. *Fuel* **2012**, 97, 710.
- [13] Aharon, M.; Elad, M. *SIAM J. Image Sci.* **2008**, 1, 228.
- [14] Aharon, M.; Elad, M.; Bruckstein, A. M. *IEEE Trans. Signal Proc.* **2006**, 54, 4311.
- [15] Liu, Q.-G.; Wang, S.-S.; Luo, J.-H. *J. Vis. Commun. Image* **2012**, 23, 182.
- [16] Karl, S.; Kjersti, E. *IEEE Trans. Signal Proc.* **2011**, 58, 2121.
- [17] Mehrdad, Y.; Thomas, B.; Mike, E. D. *IEEE Trans. Signal Proc.* **2009**, 57, 2178.

(Cheng, F.)