全光谱匹配算法在苹果分类识别中的应用

周万怀[1]，谢丽娟[1,2]，应义斌[1,2※]

（1. 浙江大学生物系统工程与食品科学学院，杭州 310058；2. 农业部设施农业装备与信息化重点实验室，杭州 310058）

摘　要：为进一步提高光谱匹配准确率，该研究对杰卡德相似性原理（jaccard similarity coefficient，JSC）进行改进并提出新的光谱相似度的计算方法。同时，对光谱进行一阶导数二值化，以保证改进后的算法适用于光谱的匹配。此外，对不同光谱分辨率对该算法的影响进行了研究。试验样本选用阿克苏红富士、山东红将军、陕西红富士和陕西金帅 4 个品种的苹果进行算法能力验证，在 2～128 cm^{-1} 之间，共 7 个不同水平的分辨率上进行比较。试验结果表明：该研究提出的算法正确分类识别率为 94.5%；研究提出算法在 8 或 16 cm^{-1} 分辨率水平下取得最佳分类识别结果。因此，基于 JSC 的全谱匹配算法在光谱数据库系统中的应用将有助于光谱查询精度的提高。

关键词：分析，算法，近红外光谱，光谱数据库系统，全谱匹配算法，杰卡德相似性原理，分类识别

doi：10.3969/j.issn.1002-6819.2013.19.035

中图分类号：S220.1　　**文献标志码**：A　　**文章编号**：1002-6819(2013)-19-0285-08

周万怀，谢丽娟，应义斌. 全光谱匹配算法在苹果分类识别中的应用[J]. 农业工程学报，2013，29(19)：285－292.
Zhou Wanhuai, Xie Lijuan, Ying Yibin. Application of full spectral matching algorithm in apple classification[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2013, 29(19): 285－292. (in Chinese with English abstract)

0　引　言

人们对近红外光谱的研究具有较为长远的历史，但直到 20 世纪 80 年代，近红外光谱技术才开始走向实际应用，这主要受益于近红外光谱学和化学计量学的结合。但是，传统的光谱信息保存形式（文件系统）不利于近红外光谱技术的快速应用和推广，主要体现在光谱信息二次利用效率低下。为了解决这一问题，国内外学者开始研究光谱数据库系统。

光谱数据库系统主要被用于光谱信息保存和维护[1-2]。此外，科学家们还尝试采用光谱数据库系统通过查询分析模式，对未知物进行分析[3-4]和实时监测[5]。对于查询分析模式而言，最关键的是光谱匹配算法。光谱匹配算法可以分为两大类：基于特征峰的光谱匹配算法[6-10]和基于全谱的光谱匹配算法[11-14]。基于特征峰的光谱匹配算法通过对不同光谱的特征峰的相似度来衡量 2 条光谱之间的相似度，进而确定 2 个样本之间的相似程度，特征峰匹配算法可以比较任意 2 条光谱数据，但是特征峰匹配算法通常具有较低的匹配精度[10]。基于全谱的光谱匹配算法则要对整条光谱的数据点进行比较，计算得到一个总体相似度，进而判断 2 条光谱之间的相似度。由此可以看出，全谱匹配算法具有较高的要求：相比较的光谱应有相同的波段范围和相同的数据点数。但全谱匹配算法通常具有较高的匹配精度。因此，在高精度匹配任务中通常采用全谱匹配算法。

已有研究中，常见的几类全谱匹配算法是：绝对差异法（absolutely distance，AD）、平方差法（square distance，SD）、相关系数法（correlation coefficient，CC）、光谱角法（spectral angle，SA）、欧式距离法（euclideandistance，ED）[11-14]。目前，这些算法已在化合物结构分析[3,15]、化学试剂识别[16]、简单混合物成分分析[5]等方面得以研究和应用。但尚无使用这些算法对复杂样品进行分析的报道。

通过对常见的全谱匹配算法原理分析发现，这些算法均直接采用光谱吸光度值作为距离或相似度计算的源数据。因此，这些方法难免受到噪声干扰。为消除噪声对光谱匹配结果的影响，本文探索一种新的光谱匹配思路：杰卡德相似性原理（jaccard similarity coefficient，JSC）是一种用来度量 2 个集合之间相互重叠程度的方法，它的一种变形可以计算 2 个二进制序列之间的相似度[17]。研究尝试将光谱数据的一阶导数二值化，然后采用 JSC 的变型基

收稿日期：2013-06-08　　修订日期：2013-08-27
基金项目：中国教育部博士点基金（20100101110135）
作者简介：周万怀（1983－），男，安徽阜阳人，博士生，主要从事光谱数据库系统及关键算法研究。杭州　浙江大学生物系统工程与食品科学学院，310058。Email：zhouwanhuai@yahoo.com.cn
※通信作者：应义斌（1964－），男，浙江宁海人，教授，博士生导师，主要从事农产品/食品品质与安全快速检测技术和智能装备方面的研究。杭州　浙江大学生物系统工程与食品科学学院，310058。
Email：yingyb@zju.edu.cn

于二进制的一阶导数进行光谱匹配。研究旨在建立一个水果近红外光谱数据库系统，探索基于查询分析模式的水果内部品质检测和分析方法。

1 材料与方法

1.1 试验样品

试验所用样本为市售苹果，包括阿克苏红富士，山东红将军，陕西红富士和陕西金帅 4 个品种，每个品种样品个数为 100 个。试验样品满足品种内不同产地的样品分类（阿克苏红富士和陕西红富士），品种间相同产地的样品分类（陕西红富士和陕西金帅）和品种间不同产地的样品分类（阿克苏红富士，山东红将军和陕西金帅）

所有苹果样品均为同一批次，经过筛选，保证样品大小均匀，无明显损伤。所有苹果样品采用湿布擦洗干净，放置在实验室环境中平衡 24 h。

1.2 仪器、光谱采集

研究采用美国热电尼高力公司生产的 Nexus 智能型傅里叶变换近红外光谱仪（美国 Thermo Electron 公司）。该光谱仪内部由内置光源、Vectra 干涉仪、检测器等部件构成。其光源为 50 W 钨卤近红外光源，该光学系统最高分辨率优于 0.09 cm^{-1}。试验选用与该仪器配套的 InGaAs 检测器（检测范围为 3 800～12 500 cm^{-1}）和智能漫反射附件，光谱采集软件使用与光谱仪配套的 Ominic6.0，光谱数据格式设置为吸光度光谱（Absorbance）。试验选用标准白板（聚四氟乙烯）的漫反射光谱作为背景。

采集光谱波段为 4 000～12 000 cm^{-1}，扫描次数为 32 次，分辨率为 2 cm^{-1}，样品到检测探头的距离为 0，每间隔 100 min 重新采集一次背景光谱。在试验准备和测试过程中，实验室温度控制在(20±1)℃范围内，相对湿度控制在 50%±3%的范围内，试验过程中关闭室内照明灯，使用黑色软质垫圈确保隔绝外界光，穿着黑色、棉质服装，消除静电和反射光影响。光谱采集前，首先将厚度约 3 mm 黑色软质垫圈固定在样品支撑架上（垫圈的圆孔和检测探头重合），以防止外界光干扰和阻止样品滚动，将果梗呈水平状摆放在垫圈上进行光谱采集，对每个样品沿赤道部位均匀选取 3 个光谱采集点进行光谱采集，数据分析时对每个样品的 3 条光谱求平均，以平均光谱作为该样品的光谱。

1.3 光谱噪声去除

一条样品光谱通常由 3 部分组成：样品信息，噪声信息和基线，如公式（1）所示。

$$f(signal)=f(sample)+f(baseline)+c \quad (1)$$

式中，$f(sample)$为样品信息，$f(baseline)$为基线，c 为噪声。通常，在特征提取之前要去除光谱中的基线和噪声信息。在本研究中，光谱采集时采用自动扣除基线方式，因此后续无需再进行去除基线的处理。在近红外光谱噪声去除方面，常用的去除光谱噪声的方法有移动平均法[18-20]，Savitzky-Golay 法[21-23]和小波变换法等[24-28]。这些算法都假设光谱中的所有数据点都含有噪声信息，因此对光谱的每一个数据点都要进行平滑处理。事实上，光谱中有些数据点的噪声含量非常低，平滑操作会将这些点的特征信息减弱，从而导致有效光谱信息的损失。为了保护这些噪声含量非常低，而信息含量非常高的数据点，研究提出一种波动频率统计法（count of fluctuation，CF）用于识别数据点噪声水平。

假设 $S(x, y)$是一条近红外光谱，x 表示波段，y 表示吸光度。CF 算法首先按照公式（2）计算 $S(x, y)$的一阶导数。

$$f_{i}.y_m=\frac{\left[S_{i}.y_m+1-S_{i}.y_m\right]}{\left[S_{i}.x_m+1-S_{i}.x_m\right]} \quad (2)$$

式中，$f_i.y_m$ 表示第 i 条光谱的第 m 个数据点的一阶导数值；$S_i.y_m$ 表示第 i 条光谱的第 m 个数据点的吸光度；$S_i.x_m$ 表示第 i 条光谱的第 m 个数据点的波段值。在此基础上，根据公式（3）和公式（4）给光谱曲线中的每个数据点赋予一个权值。

$$count_m=\sum_{m-width}^{m+width} temp= \quad (3)$$
$$1or0(1:\ f_i.y_{m-1}\times f_i.y_m>0;0:\ f_i.y_{m-1}\times f_i.y_m<0)$$

$$S_{i}.y_{m}.weight=\frac{count_m}{2\times width+1} \quad (4)$$

算法以光谱的第 m 个数据点为中心，以 $2\times width+1$ 为窗口宽度，在此范围之内统计所有的波峰和波谷个数($count_m$)，并将 $count_m$ 与窗口宽度 $2\times width+1$ 的比值 $S_i.y_m.weight$ 作为衡量 m 点噪声含量的依据。$S_i.y_m.weight$ 越大，说明噪声含量越高，反之说明噪声含量较低。根据经验，设定一个过滤阈值 T，当 $S_i.y_m.weight$ 大于 T 时，对该点进行平滑操作，否则该点不需要平滑操作。

1.4 杰卡德相似性原理及其变型

杰卡德相似性原理是一个用于计算集合之间重合度的方法[17]，原理如公式（5）所示

$$J(A,B)=\frac{|A\cap B|}{A\cup B} \quad (5)$$

式中，A 和 B 为 2 个相互比较的集合。杰卡德相似性原理有一种变型，使用该变型可以比较 2 个二进制序列之前的相似度。假设 A={0101001101…}，B={1101011101…}是 2 个长度为 n 的二进制序列，则有以下计法：

D_{11} 为 A，B 中在相同的位置同时出现 1 的次数

的统计；D_{00} 为 A，B 中在相同的位置同时出现 0 的次数的统计；D_{10} 为 A，B 中在相同的位置出现 1，0 的次数的统计；D_{01} 为 A，B 中在相同的位置出现 0，1 的次数的统计。其中，$D_{11}+D_{00}+D_{10}+D_{01}=n$，于是，杰卡德相似系数和杰卡德距离被分别定义为公式（6）和公式（7）。

$$J=\frac{D_{11}}{D_{10}+D_{01}+D_{11}} \tag{6}$$

$$J=\frac{D_{10}+D_{01}}{D_{10}+D_{01}+D_{11}} \tag{7}$$

在此算法的基础上，研究提出一种光谱匹配算法。算法原理如下：

假设 2 条光谱数据分别为 $S_1(x，y)$和 $S_2(x，y)$。分别计算 2 条光谱的一阶导数，并记为 $S'_1(x,y)$ 和 $S'_2(x,y)$。当一阶导数大于 0 时，表示曲线在相应位置为上升趋势；而当一阶导数小于 0 时，表示曲线在相应的位置为下降趋势；当一阶导数等于 0 时，表示曲线在当前位置为水平，由于一段水平曲线可以看成上升或下降趋势的延续，因此可以采用相邻的非零一阶导数值替换零值一阶导数。于是，算法第一步对一阶导数进行如下的转换

$$S'_i(x,y)=\begin{cases}S'_i(x,y) & S'_i(x,y)>0\\ S'_i(x,y) & S'_i(x,y)<0\\ S'_{i-1}(x,y) & S'_i(x,y)=0\end{cases} \tag{8}$$

以上变换实现将一阶导数的零值用相邻且非零的一阶导数值替换，从而达到消除零值的影响。接下来，将实现对一阶导数二值化。转换方法如公式（9）所示

$$S'_i(x,y)=\begin{cases}1 & S'_i(x,y)>0\\ 0 & S'_i(x,y)<0\end{cases} \tag{9}$$

通过以上的变换，得到二进制一阶导数，仍记为 $S'_1(x，y)$和 $S'_2(x，y)$。继续做以下规定：$D_{11\text{ or }00}$ 为 2 条光谱在相同位置的一阶导数同时为 1 或同时为 0 的次数；$D_{10\text{ or }01}$ 为 2 条光谱在相同位置的一阶导数不相等的次数。

满足：

$$D_{11\text{ or }00}+D_{10\text{ or }01}=n$$

其中 n 为一阶导数的数据点数。因此，2 条光谱之间的相似系数 JSC 被定义为公式（10）。

$$JSC=\frac{D_{11\text{ or }00}}{D_{11\text{ or }00}+D_{10\text{ or }01}} \tag{10}$$

2　结果与分析

2.1　分类中心构建

试验共采用 4 个品种的苹果样品近红外光谱作为算法分类识别能力的测试数据。因此，首先需要分别对 4 类样品构建类型中心。在本研究中，采用总体平均法计算类型中心。即对品种内部的所有样品光谱求均值，将该均值作为类型中心。4 类样本光谱如图 1 所示，4 个类型中心如图 2 所示。

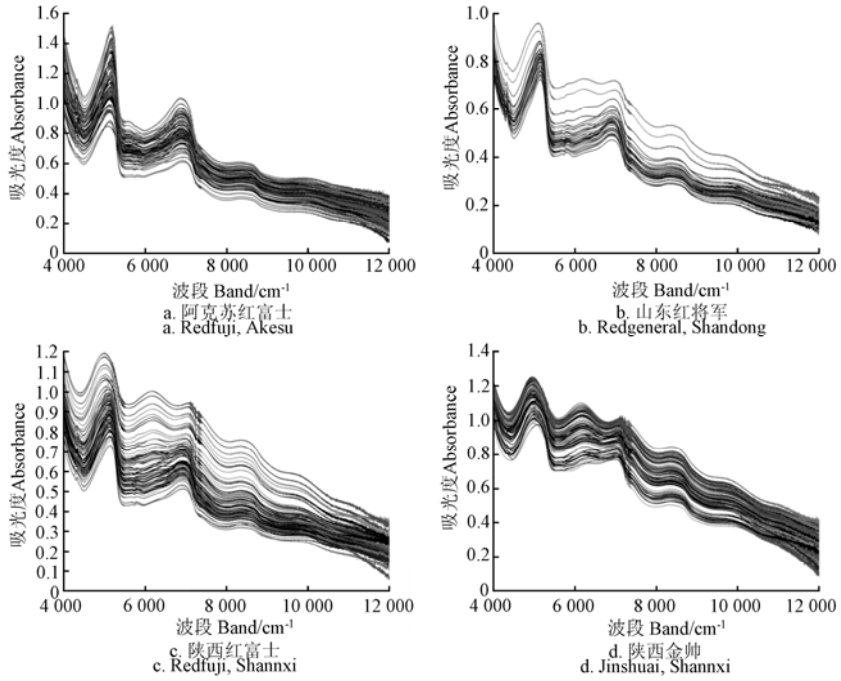


图 1　4 类苹果样品的原始光谱图

Fig.1　Raw spectra of all samples

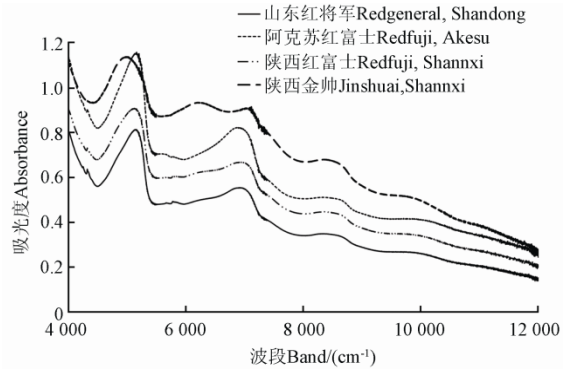


图 2 4 个类型中心

Fig.2 Four class centers

2.2 样品分类识别

对于查询和匹配算法而言，匹配精度是衡量算法性能的最为重要的指标。为了验证本文算法的性能，本研究将绝对差异法、平方差法、相关系数法、光谱角法、欧式距离法和本文提出的算法一同在系统中实现。试验对 400 个苹果样品进行分类识别，观察各个算法对各类样品的分类识别正确率，结果如表 1 所示。表 1 的统计结果表明：在 6 类匹配算法中，本文提出的匹配算法具有最高的正确分类识别结果：对 4 类样品的平均分类正确识别率达到 94%，对部分样品正确分类识别率达 100%，较其他算法具有较大的优势。

表 1 2 cm^{-1} 分辨率下的 6 种算法对 4 类样品的分类识别率

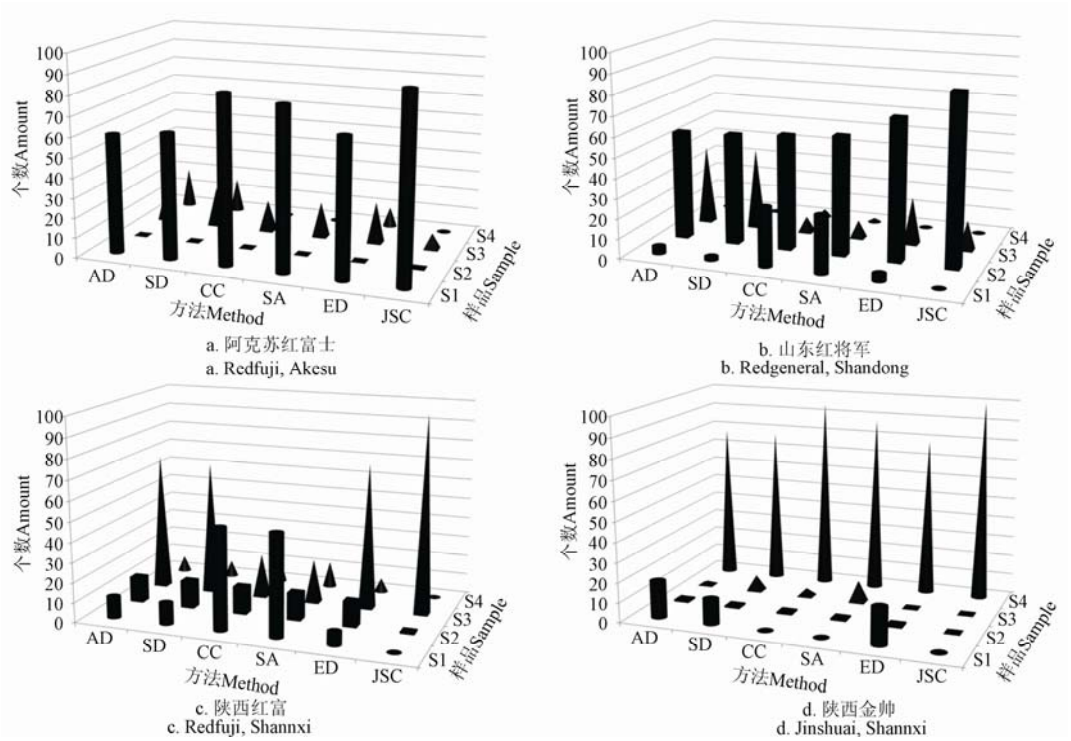
Table 1Identification results of all samples (2 cm^{-1})

样品 Sample	识别率 Identification rate/%					
	AD	SD	CC	SA	ED	JSC
S1	60	63	84	81	69	92
S2	55	56	58	60	71	85
S3	69	67	22	22	73	99
S4	78	78	95	88	79	100
平均精度 Mean	65.5	66	64.75	62.75	73	94

注：AD 为绝对差异法，SD 为平方差法，CC 为相关系数法，SA 为光谱角法，ED 为欧几里得距离法，JSC 为杰卡德相似性原理法。S1 为阿克苏红富士，S2 为山东红将军，S3 为陕西红富士，S4 为陕西金帅，Avg 为 4 类样品的平均分类正确率。下同。

Note: AD: Absolutely distance algorithm, SD: Square distance algorithm, CC: Correlation coefficient algorithm, SA: Spectral angle algorithm, ED: Euclidean distance, JSC: Jaccard similarity coefficient algorithm. S1: Redfuji produced in Akesu, S2: Redgeneral produced in Shandong, S3: Redfuji produced in Shanxi, S4: Jinshuai produced in Shanxi, AVG: Average accuracy of these four. The same as below.

进一步观察每个品种的样品对 4 个类型中心分类的结果，详细的分类识别结果如图 3 所示。可以发现：S1（阿克苏红富士）和 S3（陕西红富士）之间存在较大的误判率；S1、S2（山东红将军）、S3 和 S4（陕西金帅）存在误判率低。这些结果表明：品种内部之间的分类难度较大；同色系的样品的分类难度较大。



注：S1 为阿克苏红富士，S2 为山东红将军，S3 为陕西红富士，S4 为陕西金帅；AD 为绝对差异法，SD 为平方差法，CC 为相关系数法，SA 为光谱角法，ED 为欧几里得距离法，JSC 为杰卡德相似性原理法。

Note: S1: Redfuji produced in Akesu, S2: Redgeneral produced in Shandong, S3: Redfuji produced in Shanxi and S4: Jinshuai produced in Shanxi; AD: Absolutely distance algorithm, SD: Square distance algorithm, CC: Correlation coefficient algorithm, SA: Spectral angle algorithm; ED: Euclidean distance, JSC: Jaccard similarity coefficient algorithm.

图 3 6 种算法对 4 类样品分类识别结果

Fig.3 Classification results with six different algorithms

2.3　分辨率对算法性能的影响

根据 1.4 节中对研究所提出光谱匹配算法原理可知，杰卡德相似性原理法的工作原理是相似形理论。其比较的是 2 条光谱曲线在相同的位置是否具有相同的增减趋势，最终将具有相同增减趋势的统计数与总的光谱数据点数的比值作为 2 条光谱之间的相似性度量。可以看出，光谱分辨率对该原理具有较大的影响。因此，本文在不同分辨率水平上对算法的影响进行研究。

已有研究成果表明，水果光谱的最佳建模分辨率位于 4～16 cm^{-1} 之间[29-31]。试验采集原始光谱数据为 2 cm^{-1}，通过分辨率调整算法，分别生成了 4、8、16、32、64、128 cm^{-1} 的 6 组新的不同水平分辨率的光谱数据。鉴于分辨率为 128 cm^{-1} 时，光谱曲线线型变化已十分明显，原曲线特征信息丢失严重，没有进一步降低分辨率。在此 7 个不同分辨率水平下，研究分辨率对本文提出的算法的影响。试验结果如表 2 所示，可以看出：在 2～16 cm^{-1} 范围内，算法的分类识别正确率较为稳定，当分辨率低于 32 cm^{-1} 时，算法分类正确率呈现明显的下降趋势。

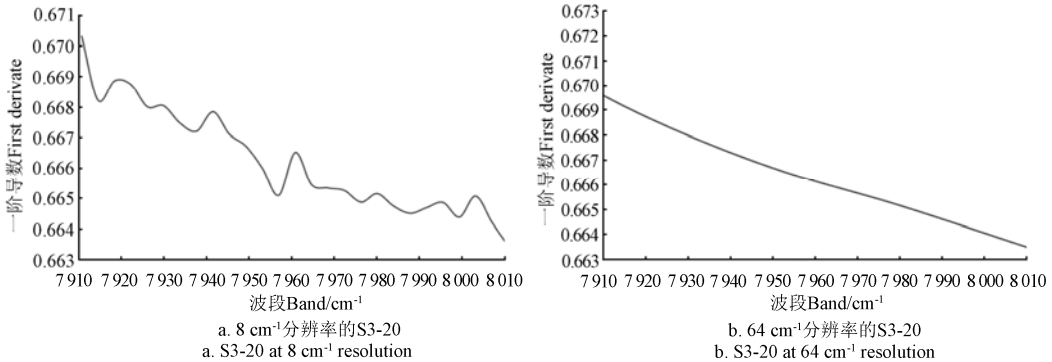
表 2　不同分辨率水平对杰卡德相似性原理法的影响

Table 2　Average identification accuracy at resolutions 2～128 cm^{-1}

样品 Sample	识别率 Identification rate/%						
	2 cm^{-1}	4 cm^{-1}	8 cm^{-1}	16 cm^{-1}	32 cm^{-1}	64 cm^{-1}	128 cm^{-1}
S1	92	92	94	94	93	90	88
S2	85	85	85	85	93	73	56
S3	99	99	99	99	84	77	70
S4	100	100	100	100	97	90	85
平均精度 Mean	94	94	94.5	94.5	91.75	83.75	76.5

为了进一步说明光谱分辨率对杰卡德相似性原理算法的影响原理，研究选择一个在 2～16 cm^{-1} 正确分类，在 32～128 cm^{-1} 范围错误分类的样品 S3-20（陕西红富士第 20 号样品），详细分析了该样品与类型中心匹配的详细过程和数据。为了观察详细的数据变换和比较细节，选取 7 910～8 010 cm^{-1} 波段，观察 S3-20 和类型中心 S3-Class 和 S4-Class 的一阶导数变换结果和 S3-20 到 2 个类型中心距离计算结果（图 4）。

从图中可以看出，8 cm^{-1} 下的 S3-20 在选取的波段内起伏不平（图 4a），但是，64 cm^{-1} 下的 S3-20 则单调递减（图 4b）。对两者的一阶导数进行二值化变换，由于在 8 cm^{-1} 分辨率下光谱起伏不平，曲线有增有减，变换后的一阶导数 0，1 相间分布；而在 64 cm^{-1} 分辨率下光谱整体单调递减，一阶导数均为负值，变换后的一阶导数全部为 0 值。对 8、64 cm^{-1} 分辨率下的 S3-Class 和 S4-Class 光谱的一阶导数进行二值化变换。用 JSC 算法计算 S3-20 与 S3-Class 和 S4-Class 的匹配度：在 8 cm^{-1} 下，S3-20 与 S3-Class 的一阶导数值在相同波相等的次数为 26，在此波段内总的数据点数亦为 26，所以 S3-20 与 S3-Class 的相似度为 26/26；同理 S3-20 与 S4-class 的相似度为 10/26，S3-20 与 S3-Class 的匹配度较高，综合考虑整条光谱，S3-20 与 S3-Class 的相似度大于 S3-20 与 S4-Class 的相似度，所以 S3-20 被正确分类到 S3-Class 中；在 64 cm^{-1} 下，S3-20 与 S3-Class 的相似度为 3/3，而 S3-20 与 S4-Class 的相似度也为 3/3，综合考虑整条光谱，S3-20 与 S3-Class 的相似度小于 S3-20 与 S4-Class 的相似度，因此误分到 S4-Class 中。



a. 8 cm^{-1}分辨率的S3-20
a. S3-20 at 8 cm^{-1} resolution

b. 64 cm^{-1}分辨率的S3-20
b. S3-20 at 64 cm^{-1} resolution

注：S3-20 是陕西红富士第 20 号样本。
Note: S3-20: the 20th sample in S3.

图 4　不同分辨率导致不同分类识别结果示意图

Fig.4　Spectral resolution effect on classification result

3　结　论

本研究针对水果近红外光谱数据库系统开发需求，在杰卡德相似性原理的基础上，提出一种改进的光谱匹配算法。试验采用 400 个苹果样品（4 个品种，每个品种 100 个）对算法能力进

行验证，结果表明 JSC 算法对所有样品的总体正确分类识别率为 94.5%，对部分样品正确分类识别率高达 100%。其他算法对所有样品总体正确分类识别率最高为 73%。此外，研究了光谱分辨率对 JSC 算法的影响，发现 JSC 算法对光谱分辨率较为敏感，且其相对较优分类识别分辨率为 8 或 16 cm^{-1}。根据研究结果，本研究提出光谱匹配算法可用于复杂成分样品的分类识别，对于苹果近红外光谱而言，算法理想工作分辨率为 8 或 16 cm^{-1}，当分辨率高于 8 cm^{-1}，由于光谱噪声含量过高导致分类结果略有降低，而当分辨率低于 16 cm^{-1} 时，样品信息损失严重，导致分类正确率明显下降。由于算法对分辨率敏感，用于不同类型的样品时，首先需要找出算法对该类样品的最佳工作分辨率。

[参 考 文 献]

[1] Penchev P N, Miteva V L, Sohou A N, et al. Implementation and testing of routine procedure for mixture analysis by search in infrared spectral library[J]. Bulgarian Chemical Communications, 2008, 40(4): 1－5.

[2] Johnson T J, Profeta L T M, Sams R L, et al. An infrared spectral database for detection of gases emitted by biomass burning[J]. Vibrational Spectroscopy, 2010, 53(1): 97－102.

[3] Deübska B, Guzowska-Swider B, Cabrol-Bass D. Automatic generaten of knowledge base from infrared spectral database for substructure identification[J]. Journal of Chemical information and Modeling, 2000, 4(2): 330－338.

[4] Karpushkin E, Bogomolov A, Zhukov Y, et al. New system for computer-aided infrared and Raman spectrum interpretation[J]. Chemometrics and Intelligent Laboratory Systems, 2007, 88(1): 107－117.

[5] Chu P M, Rhoderick G C, Vlack D V, et al. A quantitative infrared spectral database of hazardous air pollutants[J]. Fresenius' Journal of Analytical Chemistry, 1998, 360(3/4): 426－429.

[6] Coombes K R, Fritsche H A, Clarke C, et al. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization[J]. Clinical Chemistry, 2003, 49: 1615－1623.

[7] Lau O W, Hon P K, Bai T. A new approach to a coding and retrieval system for infrared spectral data: The effective peaks matching's method[J]. Vibrational Spectroscopy, 2000, 23(1): 23－30.

[8] Vivó-Truyols G, Torres-Lapasió J R, Van Nederkassel A M, et al. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part I: Peak detection[J]. Journal of Chromatography A, 2005, 1096(1): 133－145.

[9] Vivó-Truyols G, Torres-Lapasió J R, Van Nederkassel A M, et al. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part II: Peak model and deconvolution algorithms[J]. Journal of Chromatography A, 2005, 1096(1): 146－155.

[10] Yang Chao, He Zengyou, Yu Weichuan. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis[J]. BMC Bioinformatics, 2009, 10(4): 1－13.

[11] Li Jianfeng, Hibbert D B, Fuller S, et al. A comparative study of point-to-point algorithms for matching spectra[J]. Chemometrics and Intelligent Laboratory Systems, 2006, 82(1/2): 50－58.

[12] Loudmilk J B, Himmelsbach D S, Barton F E, et al. Novel search algorithms for a mid-infrared spectral library of cotton contaminants[J]. Applied Spectroscopy, 2008, 62(6): 661－670.

[13] Leung A K, Chau F, Gao J, et al. Application of wavelettransform in infrared spectrometry: Spectral compression and library search[J]. Chemometrics and Intelligent Laboratory Systems, 1998, 43(1/2): 69－88.

[14] Penchev P N, Sohou A N, Andreev G N. Description and performance analysis of an Infrared library search system[J]. Spectroscopy letters, 1996, 29(8): 1513－1522.

[15] Varmuza K, Penchev P N, Scsibrany H. Large and frequently occurring substructures in organic comppunds obtained by library search of infrared spectra[J]. Vibrational Spectroscopy, 1999, 19(2): 407－412.

[16] Yoon W L, Jee R D, Moffat A C. An interlaboratory trial to study the transfer ability of a spectral library for the identification of solvents using near-infrared spectroscopy[J]. Analyst, 2000, 125(10): 1817－1822.

[17] Varmuza K, Karlovits M, Demuth W. Spectral similarity versus structural similarity: infrared spectroscopy[J]. Analytica Chimica Acta, 2003, 490(1/2): 313－324.

[18] 严衍禄，陈斌，朱大洲，等. 近红外光谱分析原理、技术与应用[M]. 北京：中国轻工业出版社，2005.

[19] 万鹏，朱洁，陈贻范. 移动平均法的数字滤波特性分析[J]. 海洋技术，1997，16(3)：29－31.
Wan Peng, Zhu Jie, Chen Yifan. The characteristic

analysis of digital filter about the average over moving[J]. Ocean Technology, 1997, 16(3): 29－31. (in Chinese with English abstract)

[20] 孙旭东，郝勇，高荣杰，等. 脐橙糖度近红外光谱在线检测数学模型优化研究[J]. 光谱学与光谱分析，2011，31(5)：1230－1235.
Sun Xudong, Hao Yong, Gao Rongjie, et al. Research on optimization of model for detecting sugar content of navel orange by online near infrared spectroscopy[J]. Spectroscopy and spectral analysis, 2011, 31(5): 1230－1235. (in Chinese with English abstract)

[21] 谢军，潘涛，陈洁梅，等. 血糖近红外光谱分析的 Savitzky-Golay 平滑模式与偏最小二乘法因子数的联合优选[J]. 分析化学，2010，38(3)：342－346.
Xie Jun, Pan Tao, Chen Jiemei, et al. Joint optimization of Savitzky-Golay smoothing models and partial least squares factors for near-infrared spectroscopic analysis of serum glucose[J]. Chinese journal of analytical chemistry, 2010, 38(3): 342 － 346. (in Chinese with English abstract)

[22] 孙通，应义斌，刘魁武，等. 梨可溶性固形物含量的在线近红外光谱检测[J]. 光谱学与光谱分析，2008，28(11)：2536－2539.
Sun Tong, Ying Yibin, Liu Kuiwu, et al. Online detection of soluble solids content of pear by near infrared transmission spectrum[J]. Spectroscopy and Spectral Analysis, 2008, 28(11): 2536－2539. (in Chinese with English abstract)

[23] 陈华舟，潘涛，陈洁梅. 多元散射校正与 Savitzky-Golay 平滑模式的组合优选应用于土壤有机质的近红外光谱分析[J]. 计算机与应用化学，2011，28(5)：518－522.
Chen Huazhou, Pan Tao, Chen Jiemei. Combination optimization of multiple scatter correction and Savitzky-Golay smoothing modes applied to the near infrared spectroscopy analysis of soil organic matter[J]. Computers and applied chemistry, 2011, 28(5): 518－522. (in Chinese with English abstract)

[24] 田高友，袁洪福，刘慧颖，等. 小波变换在近红外光谱分析中的应用进展[J]. 光谱学与光谱分析，2003，23(6)：1111－1114.
Tian Gaoyou, Yuan Hongfu, Liu Huiying, et al. The application of wavelet transform in near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2003, 23(6): 1111－1114. (in Chinese with English abstract)

[25] 田高友，袁洪福，刘慧颖，等. 小波变换用于近红外光谱性质分析[J]. 分析化学，2004，32(9)：1125－1130.
Tian Gaoyou, Yuan Hongfu, Liu Huiying, et al. Application of wavelet transform on the near infrared analysis[J]. Chinese journal of Analytical Chemistry, 2004, 32(9): 1125－1130. (in Chinese with English abstract)

[26] 褚小立，袁洪福，陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展，2004，16(4)：528－542.
Chu Xiaoli, Yuan Hongfu, Lu Wanzhen. Progress and application of spectral data pretreatment and wavelet selection methods in NIR analytical technique[J]. Progress in Chemistry, 2004, 16(4): 528－542. (in Chinese with English abstract)

[27] 应义斌，刘燕德，傅霞萍. 基于小波变换的水果糖度近红外光谱检测研究[J]. 光谱学与光谱分析，2006，26(1):63-66.
Ying Yibin, Liu Yande, Fu Xiaping. Sugar content prediction of apple using near infrared spectroscopy treated by wavelet transform[J]. Spectroscopy and Spectral Analysis, 2006, 26(1): 63－66. (in Chinese with English abstract)

[28] 郝勇，陈斌，朱锐. 近红外光谱预处理中几种小波消噪方法的分析[J]. 光谱学与光谱分析，2006，26(10)：1838－1841.
Hao Yong, Chen Bin, Zhu Rui. Analysis of several methods for wavelet denoising in near infrared spectrum pretreatment[J]. Spectroscopy and Spectral Analysis, 2006, 26(10): 1838－1841. (in Chinese with English abstract)

[29] Rodriguez-Saona L E, Fry F S, McLaughlin M A, et al. Rapid analysis of sugars in fruit juices by FT-NIR spectroscopy[J]. Carbohydrate Research, 2001, 336(1): 63－74.

[30] Liu Yande, Ying Yibin. Use of FT-NIR spectrometry in non-invasive measurements of internal quality of ‘Fuji’ apples[J]. Postharvest Biology and Technology, 2005, 37(1): 65－71.

[31] 谢丽娟，刘东红，张宇环，等. 分辨率对近红外光谱和定量分析的影响研究[J]. 光谱学与光谱分析，2007，27(8)：1489－1492.
Xie Lijuan, Liu Donghong, Zhang Yuhuan, et al. Study on the influence of resolution on near infrared spectra and quantitative analysis[J]. Spectroscopy and Spectral Analysis, 2007, 27(8): 1489－1492. (in Chinese with English abstract)

Application of full spectral matching algorithm in apple classification

Zhou Wanhuai[1], Xie Lijuan[1,2], Ying Yibin[1,2※]

(1. *School of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou* 310058, *China;* 2. *Key Laboratory of Equipment and Informatization in Environment Controlled Agriculture, Ministry of Agriculture, Hangzhou* 310058, *China*)

Abstract: A spectral database system (SDBS) can improve the usage efficiency and expand the application scope of spectra and their feature information, mainly referring to spectral peak information. The spectral matching algorithm (SMA) plays a decisive role in SDBS for the SMA which determines the similarity between the sample spectrum and reference spectrum, and further, decides the accuracy of database query. Traditional full spectral matching algorithms compute the distance or similarity among different spectra with spectral absorbance or reflectance directly, so they are vulnerable to noise. For a higher accuracy of a full spectral matching algorithm, this paper presents a full spectral matching algorithm based on a Jaccard similarity coefficient (JSC). JSC is a useful measure of the overlap that *A* and *B* have the same attributes which should either be 0 or 1. In order to satisfy the requirement of JSC, the first derivate of raw spectra should be computed, and a transformation process would transform negative values (of the first-order derivate) to 0 and positive values to 1, where 0 means the raw spectrum is descending in the according small region while 1 means the raw spectrum is ascending in the according small region. Different from common full spectral matching algorithms, the new proposed one calculates the similarity between different spectra with a spectral waveform but not with the absorbance or reflectance directly. Therefore, the influence of absolute absorbance or reflectance intensity was reduced and the influence of the similarity of the spectral waveform was enhanced. This mean that what substances are contained in the sample is more important than the contents of these substances. In this way, the influence of noise and the differences caused by different spectral collecting areas of solid samples was reduced to a quite low level. Comparisons among common full spectral matching algorithms and our new proposed algorithm have been carried out, and the results showed that 94.5% of the samples were correctly classified by our new proposed algorithm (4 varieties of apples, each number was 100) and the second highest classification accuracy was 73% obtained with a Euclidean distance (ED) method. This conclusion indicated that the proposed algorithm was more suitable for the classification of different kinds of samples and it would be helpful to reduce the database query scope, shorten the time consuming, and improve the accuracy of the data query. From the principle of this algorithm, it was obvious that it must be affected by the interval among the data points of the spectra. Thus, the effect of spectral resolution on the proposed algorithm was studied. In total, seven different resolutions (2~128 cm^{-1}) were tested. It is a pity that our new proposed algorithm is sensitive to spectral resolution and the optimal resolution for this algorithm approximately is 8 or 16 cm^{-1} for apples' near infrared spectra. Therefore, the optimal resolution of this algorithm should be determined at first when it is used for the spectral matching of new objects. In short, our proposed spectral matching algorithm can classify NIR spectra of solid samples with higher accuracy and the application of this algorithm will be helpful in improving the accuracy of a spectral database query.

Key words: spectral analysis, algorithms, near infrared spectroscopy, spectra database system, full spectra matching algorithm, jaccard similarity coefficient, classification

（责任编辑：刘丽英）