DOI:10.6041/j.issn.1000-1298.2012.10.021

基于 OC - SVM 和近红外光谱的秸秆固态发酵进程监测*

江　辉　刘国海　梅从立　肖夏宏　于　霜　丁煜函
(江苏大学电气信息工程学院,镇江 212013)

【摘要】 利用近红外光谱技术结合一类支持向量机(OC - SVM)快速监测秸秆蛋白饲料固态发酵进程。首先获取发酵物样本在 10 000 ~ 4 000 cm^{-1}波数范围内的近红外漫反射光谱并对其进行主成分分析,提取前 7 个主成分因子作为模型的输入变量,然后运用 OC - SVM 算法建立判别模型。在模型建立过程中,采用交互验证的方法优化 OC - SVM 模型的相关参数。实验结果表明,在相同的条件下,OC - SVM 模型在处理失衡训练样本的问题上明显优于 SVM 模型,当训练集中目标类和非目标类样本数比为 1∶8时,OC - SVM 模型在验证集中的正确判别率达到 85%。

关键词:秸秆　固态发酵　近红外光谱　主成分分析　一类支持向量机
中图分类号:O657.33;TQ92　　**文献标识码**:A　　**文章编号**:1000-1298(2012)10-0114-04

Monitoring of Straw Solid-state Fermentation Based on NIR and One-class Support Vector Machine

Jiang Hui　Liu Guohai　Mei Congli　Xiao Xiahong　Yu Shuang　Ding Yuhan
(*School of Electronic and Information Engineering*, *Jiangsu University*, *Zhenjiang* 212013, *China*)

Abstract

Near infrared (NIR) spectroscopy coupled with one-class support vector machine (OC - SVM) were used to rapidly and accurately monitor physical and chemical changes in solid-state fermentation (SSF) of crop straws without the need for chemical analysis. Raw spectra of fermented samples were acquired with wavelength range of 10 000 ~ 4 000 cm^{-1}. Then the top seven PCs as input vectors were extracted by principal component analysis (PCA). OC - SVM algorithm was implemented to develop identification model, and some parameters of OC - SVM model were optimized by cross-validation in calibrating model. Experimental results showed that OC - SVM model revealed its incomparable superiority than SVM model in handling imbalance training sets under the same condition. The discrimination rate of OC - SVM model was 85% in the validation set when the ratio of samples from target class to those from non-target class was one to eight in the training set.

Key words　Straw, Solid-state fermentation, Near infrared spectroscopy, Principal component analysis, One-class support vector machine

引言

我国作为农业大国,每年可产农作物秸秆达 7 亿t 之多,其中用于加工成饲料的不到 15%,大部分秸秆被直接还田或焚烧,焚烧造成的资源浪费和环境污染已引起全社会的关注[1]。随着农业生物技术的发展,利用微生物发酵手段来处理农作物秸秆,使其成为家畜饲料,既能促进我国畜牧业的发

收稿日期:2011 - 10 - 30　修回日期:2011 - 11 - 10
* 国家高技术研究发展计划(863 计划)资助项目(2007AA04Z179)、江苏省研究生科研创新计划资助项目(CXZZ11_0572)、江苏高校优势学科建设工程资助项目(PAPD(2011)6)和镇江市农业科技支撑资助项目(NY2010017)
作者简介:江辉,博士生,主要从事软测量技术研究,E-mail:jiangh1118@yahoo.cn
通讯作者:刘国海,教授,博士生导师,主要从事软测量控制研究,E-mail:ghliu@ujs.edu.cn

展，又能改善农业生态环境，实现资源的充分利用。近红外光谱（near infrared spectroscopy，简称 NIR）分析技术因其速度快、成本低且重现性好等优点，已在众多领域得到了广泛应用[2~5]，同时也被应用于液态发酵过程的监控和在线分析[6~10]，但是目前 NIR 分析技术在固态发酵过程中的应用研究还很少。

一类支持向量机（one-class support vector machine，简称 OC－SVM）是在传统二分类支持向量机的基础上衍生的一种新的分类方法[11]，其目的是将目标类样本与非目标类样本进行有效区分，具有无需先验知识和结构风险最小化等优点，因此被广泛应用于故障诊断[12]和异常检测[13]中。鉴于此，本研究提出基于 NIR 和 OC－SVM 结合的监测秸秆蛋白饲料固态发酵进程，以便更好地控制整个固态发酵过程。为了体现 OC－SVM 算法在处理失衡训练样本问题上的优势，将其结果与传统二分类 SVM 模型的结果进行比较。

1　材料与方法

1.1　样本准备及划分

在小麦秸秆固态发酵生产蛋白饲料过程中，利用 GTG－100 型固态发酵装置，每间隔 12 h 采集 4 个样本，整个发酵进程（0～72 h，共 7 个采样点）可采集发酵产物样本 28 个。利用相同的材料及条件发酵 5 个批次，共收集发酵样本 140 个。

通常情况下，在整个发酵进程中，随着底物消耗和微生物增长，固态基质的 pH 值是逐渐减小的。图 1 所示为发酵时间与基质 pH 值之间的关系。从图中可以看出，在发酵起始 12 h 和发酵 60 h 以后，基质 pH 值的变化甚微，而在 12～60 h 之间，pH 值的下降趋势几乎呈一条直线。因此，基于 pH 值的实测数据，研究将整个固态发酵进程随着时间推移划分为 3 个阶段，即延滞期（采样点为 0 和 12 h，共 40 个样本）、指数期（采样点为 24、36 和 48 h，共 60 个样本）和稳定期（采样点为 60 和 72 h，共 40 个样本）。

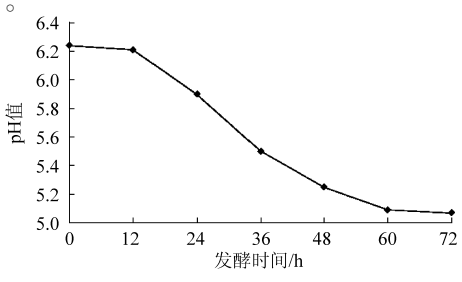

图 1　发酵时间与基质 pH 值的关系

Fig. 1　Relation between incubation time and pH value of fermented substrate

在研究建立 OC－SVM 判别模型时，将稳定期阶段的发酵样本视为目标类（40 个样本），延滞期和指数期阶段的发酵样本划分为非目标类（100 个样本）。为了显示出 OC－SVM 在处理失衡训练样本问题上的优势，本研究将训练集和验证集的样本划分如下：训练集包含目标类样本 10 个，非目标类样本分别有 10、20、40 和 80 个，使得训练集中目标类和非目标类样本数比分别为 1∶1、1∶2、1∶4 和 1∶8；验证集中目标类和非目标样本各有 10 个。

1.2　光谱数据采集

利用 Antaris Ⅱ 傅里叶变换近红外光谱仪（Thermo Scientific Ltd.，USA）采集发酵样本的近红外光谱。扫描波数范围为 10 000～4 000 cm^{-1}，扫描次数为 16 次，分辨率为 8 cm^{-1}，每条光谱包含 1 557 个数据点。实验时，室内温度保持在 25℃ 左右，湿度恒定。将样本放入标准样品杯中，充分压实。每个样本在不同位置采集 3 次，并将其平均光谱作为该样本的原始光谱。

图 2 所示为发酵样本在延滞期、指数期和稳定期阶段的近红外光谱。从图 2 中可以看出，不同发酵阶段样本光谱数据间的差异还是比较明显的。在整个固态发酵过程中，随着底物的消耗、微生物的增长及目标产物的合成，固态基质的有机物成分发生了巨大变化，而这些有机物的含氢基团（如 C—H、O—H、S—H 和 N—H 等）能在近红外区域形成倍频和合频的吸收，如果发酵样本内部有机物成分或含量发生了变化，那么它们就能在近红外光谱上表现出不同强度的吸收信号。因此，本研究正是基于这一原理，利用 NIR 结合 OC－SVM 模式识别方法快速监测秸秆固态发酵的进程。

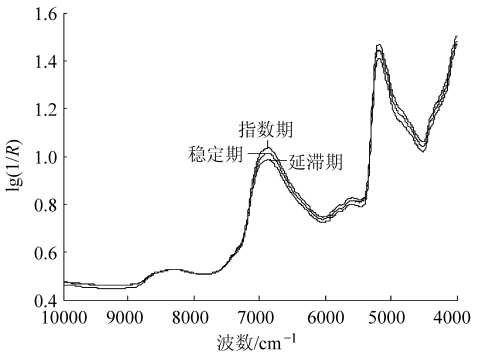

图 2　发酵样本在不同阶段的近红外光谱

Fig. 2　Spectra of fermented samples in different phases

1.3　数据分析方法

一类支持向量机[11]的基本原理是：对于给定的数据集 $X=\{x_i \mid x_i \in \mathbf{R}^d\}_{i=1}^n$，通过核函数 Φ 将样本数据映射到高维特征空间中，然后再在该空间中找到一个体积最小且尽可能多地包含目标类样本，使

其与非目标类样本以最大间隔分开。该问题可以通过二次规划问题来求解,即

$$\begin{cases}\min & \frac{1}{2}\|w\|_2^2+\frac{1}{vn}\sum_{i=1}^{n}\xi_i-b\\ \text{s.t.} & \langle w,\Phi(x_i)\rangle\geqslant b-\xi_i \quad (\xi_i\geqslant 0)\end{cases} \tag{1}$$

式中 n——训练样本的数量

v——平衡参数 ξ——松弛变量

w——权值 b——阈值

通过引入拉格朗日乘子来解决上面的最优化问题,化简可得到该优化问题的对偶形式为

$$\begin{cases}\min & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jK(x_i,x_j)\\ \text{s.t.} & 0\leqslant a_i\leqslant\frac{1}{vn} \quad \sum_{i=1}^{n}a_i=1\end{cases} \tag{2}$$

式中 $K(x_i, x_j)$——核函数

a_i——Lagrange 乘数

最小化方程(2),可得到一系列的 a_i,其中大多数的 a_i为零,只有少部分大于零,其对应的样本 x_i称为支持向量,决策函数定义为

$$f(x)=\sum_{i=1}^{n}\alpha_iK(x_i,x)-b \tag{3}$$

当 $f(x)>0$ 时,x 位于该分布区域之内;$f(x)=0$ 时,x 位于该分布区域边界上;$f(x)<0$ 时,x 位于该分布区域之外。对于待测样本,依据其落在特征空间中的哪一侧来判别其所属类别。

本研究所有数据分析是在 Matlab R2010a (Mathworks, USA)软件平台上完成的。

2 结果与讨论

2.1 主成分分析

由于发酵样本中许多有机物的含氢基团都能在近红外区域形成倍频和合频的吸收。因此,获取的光谱数据间存在严重的相关性,造成大量信息冗余。主成分分析[14](principal component analysis, 简称 PCA)是沿着协方差最大方向将高维光谱数据空间向低维数据空间投影,且各主成分得分向量间相互正交,把多个指标化为几个综合指标的一种统计方法。通过 PCA,可以达到简化数据、消除冗余信息、以提高模型性能的目的。图 3 为不同发酵阶段所有光谱数据矩阵前 3 个主成分的得分图。从图 3中可以方便地观察各个发酵阶段样本点在 3 维空间中的分布情况,尽管不能通过主成分分析将其直接区分开,但它们的聚类趋势是很明显的。通过分析发现,前 7 个主成分因子的累计方差贡献率达到 99.31%,这 7 个主成分因子几乎能完全反映原始光谱数据的信息。因此,本研究采用前 7 个主成分因子作为 OC-SVM 判别模型的输入变量。

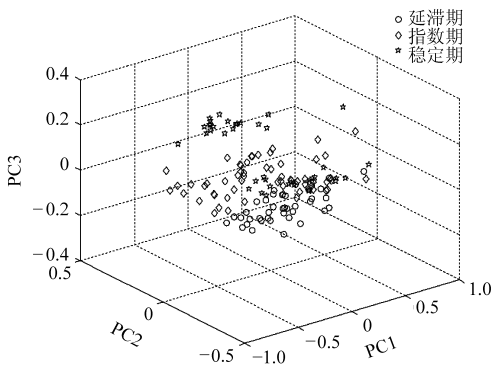

图 3 发酵样本的前 3 个主成分得分图

Fig. 3 Score cluster plot with the top three PCs for all fermented samples

2.2 OC-SVM 模型建立

OC-SVM 是利用核函数将输入向量映射到高维特征空间中,并在该空间中寻求最优分类面。目前,常见的核函数有多项式、径向基和 sigmoid 函数 3 种类型,核函数的选取对所建模型的性能影响较大。通过对 3 种核函数的比较,径向基是非线性的核函数,能较好处理样本数据间存在非线性关系的影响,且函数结构简洁,可有效降低模型训练的复杂度[15]。因此,本研究采用径向基函数作为 OC-SVM 的内核。

核函数确定后,平衡参数 v 和核函数宽度参数 σ 对所建模型的判别结果有一定的影响。因此,在模型建立过程中,有必要对这些参数进行优化,以提高预测模型的正确判别率(正确判别率是模型正确识别样本数与总样本数的比值),本研究采用网格搜索交互验证的方法优化这两个参数。图 4 为目标类与非目标类训练样本数比为 1∶8时,OC-SVM 模型参数优化结果。具体步骤为:首先,对参数 v 和 σ 在较大取值范围内进行搜索优化,结果显示,参数 v 取值范围在 0.1~1.0 之间,σ 取值范围在 1~10 之间时,OC-SVM 模型可取得较理想的判别结果;然后再在该范围内对这 2 个参数进行进一步优化,此时,研究分别选取 10 个 v(0.1~1.0,步长为 0.1)和 10 个 σ(1~10,步长为 1),并以验证集中独立样本的判别率来确定参数 v 和 σ 的最优值。从图 4 中可以看出,当 $v=0.5$、$\sigma=4$ 时,OC-SVM 模型可获得最佳判别结果,总正确判别率为 85%。

2.3 结果比较与分析

为了显示出 OC-SVM 模型在处理失衡训练样本问题上的优势,将其结果与传统二分类的 SVM 模型结果进行了比较。在 SVM 模型建立过程中,采取同样的方法对其参数进行优化。最佳 OC-SVM 和 SVM 模型结果比较如表 1 所示。从表 1 中可以看

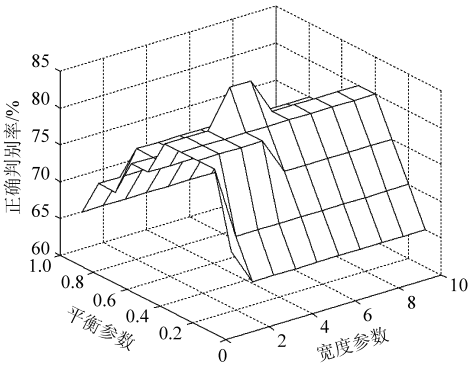

图 4　OC－SVM 模型在不同 v 和 σ 下的正确判别率

Fig. 4　Discrimination rates of OC－SVM model with different v and σ

出，当训练集中的目标类和非目标类样本数均衡时，2 种类型的 SVM 模型都能在验证集中取得较好的判别结果，随着目标类和非目标类训练样本数比例的增加，传统 SVM 模型的判别率逐渐降低，尤其是当目标类和非目标类训练样本数比达到 1∶8 时，该模型将验证集中目标类样本几乎全部判入非目标类，而在同样情况下，OC－SVM 模型的正确判别率仍然能达到 85%。

表 1　OC－SVM 和 SVM 模型结果比较

Tab. 1　Results comparisons of OC－SVM and SVM models

训练样本数比	SVM			OC－SVM		
	目标类样本数/个	非目标类样本数/个	正确判别率/%	目标类样本数/个	非目标类样本数/个	正确判别率/%
1∶1	10/10	10/10	100	9/10	9/10	90
1∶2	7/10*	10/10	85	8/10	10/10	90
1∶4	5/10	8/10	65	9/10	10/10	95
1∶8	1/10	10/10	55	8/10	9/10	85

注：* 7/10 表示共有 10 个样本，正确识别 7 个，其他类推。

传统二分类 SVM 模型在训练时需要 2 类数据。它先通过核函数将输入变量映射到高维特征空间中，然后再在该特征空间中寻求最优分类面，且最优超平面的获取取决于超平面两侧附近的样本点（即支持向量）。当训练集中的 2 类样本数量相当时，可获得较理想的判别结果；当训练集中的某一类样本数量缺乏，不足以对划分最优超平面提供足够的支持时，就会导致最优分类面出现严重偏离，过于靠近少数类别。在本研究中，当训练集中目标类与非目标类训练样本个数出现较严重的失衡时，分类结果严重向非目标类倾斜，极易造成对目标类样本的误判。

OC－SVM 算法是一种一类分类方法，主要关注目标类样本数据的特点，在模型预测时仅判别待测样本是否属于目标类，因而并不存在训练集中 2 类样本失衡的问题，且模型判别结果也不会出现向多数类倾斜。尽管 OC－SVM 主要是依靠一类（目标类）样本训练模型，但对于二分类的问题，正类或负类都可以视作目标类。因此，在实际应用中，OC－SVM 对于训练样本失衡问题的意义在于，如果一类目标判别效果不佳时，可以试图改变目标类，以期改善判别性能，从根本上避免了不同类别的训练样本存在失衡的影响，使其在处理训练样本失衡问题上比传统二分类 SVM 具有明显的优越性。

3　结束语

利用近红外光谱技术结合 OC－SVM 算法建立一类判别模型对秸秆固态发酵进程进行快速监测。相比于传统的 SVM 模型，OC－SVM 模型在处理训练集中样本失衡问题上显示出其独特的优越性，当训练集中目标类与非目标类样本数比为 1∶8 时，OC－SVM 模型在验证集中的正确判别率可达到 85%。实验结果表明利用近红外光谱分析技术快速监测固态发酵进程是可行的。

参　考　文　献

1　韩鲁佳，闫巧娟，刘向阳，等. 中国农作物秸秆资源及其利用现状[J]. 农业工程学报，2002，18(3)：87～91.
Han Lujia，Yan Qiaojuan，Liu Xiangyang，et al. Straw resources and their utilization in China[J]. Transactions of the CSAE，2002，18(3)：87～91.（in Chinese）

2　朱伟兴，江辉，陈全胜，等. 梨可溶性固形物含量 NIR 与变量筛选无损检测[J]. 农业机械学报，2010，41(10)：129～133.
Zhu Weixing，Jiang Hui，Chen Quansheng，et al. Measurement of soluble solids content in pear by FT－NIR spectroscopy and variable selection [J]. Transactions of the Chinese Society for Agricultural Machinery，2010，41(10)：129～133.（in Chinese）

3　Balabin R M，Safieva R Z. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data [J]. Fuel，2008，87(7)：1 096～1 101.

4　Chen Q，Zhao J，Liu M，et al. Determination of total polyphenols content in green tea using FT－NIR spectroscopy and different PLS algorithms [J]. Journal of Pharmaceutical and Biomedical Analysis，2008，46(3)：568～573.

（下转第 166 页）

Fan Yingli. AGV navigation identifier identification method based on machine vision[J]. Servo Control, 2011(2): 57 ~ 59. (in Chinese)

6 李进, 陈无畏, 李碧春, 等. 自动导引车视觉导航的路径识别和跟踪控制[J]. 农业机械学报, 2008, 39(2): 20 ~ 24.
Li Jin, Chen Wuwei, Li Bichun, et al. Road recognition and tracking control of an vision guided AGV[J]. Transactions of the Chinese Society for Agricultural Machinery, 2008, 39(2): 20 ~ 24. (in Chinese)

7 安秋, 李志臣, 姬长英, 等. 基于光照无关图的农业机器人视觉导航算法[J]. 农业工程学报, 2009, 25(11): 208 ~ 212.
An Qiu, Li Zhichen, Ji Changying, et al. Agricultural robot vision navigation algorithm based on illumination invariant image [J]. Transactions of the CSAE, 2009, 25(11): 208 ~ 212. (in Chinese)

8 Marchant J A, Onyango C M. Shadow-invariant classification for scenes illuminated by daylight[J]. Journal of the Optical Society of America, 2000, 17(11): 1 952 ~ 1 961.

9 丁幼春, 王书茂, 陈度. 基于图像旋转投影的导航路径检测算法[J]. 农业机械学报, 2009, 40(8): 63 ~ 66.
Ding Youchun, Wang Shumao, Chen Du. Navigation line detection arithmetic based on image rotation and projection[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(8): 63 ~ 66. (in Chinese)

10 赵博, 毛恩荣, 毛文华, 等. 农业车辆杂草环境下视觉导航路径识别方法[J]. 农业机械学报, 2009, 40 (增刊): 183 ~ 186.
Zhao Bo, Mao Enrong, Mao Wenhua, et al. Path recognition for vision navigation system of agricultural vehicle in weed environment[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(Supp.): 183 ~ 186. (in Chinese)

11 丁幼春, 廖庆喜, 黄海东, 等. 联合收获机大曲率路径视觉导航方法[J]. 农业机械学报, 2011, 42(增刊): 122 ~ 127.
Ding Youchun, Liao Qingxi, Huang Haidong, et al. Large curvature path detection for combine harvester based on vision navigation[J]. Transactions of the Chinese Society for Agricultural Machinery, 2011, 42(Supp.): 122 ~ 127. (in Chinese)

12 丁幼春, 王书茂. 联合收获机视觉导航控制系统设计与试验[J]. 农业机械学报, 2010, 41(5): 137 ~ 143.
Ding Youchun, Wang Shumao. Vision navigation control system for combine harvester[J]. Transactions of the Chinese Society for Agricultural Machinery, 2010, 41(5): 137 ~ 143. (in Chinese)

(上接第 117 页)

5 Candolfi A, De Maesschalck R, Massart D L, et al. Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA [J]. Journal of Pharmaceutical and Biomedical Analysis, 1999, 19(6): 923 ~ 935.

6 McLeod G, Clelland K, Tapp H, et al. A comparison of variate pre-selection methods for use in partial least squares regression: a case study on NIR spectroscopy applied to monitoring beer fermentation [J]. Journal of Food Engineering, 2009, 90(2): 300 ~ 307.

7 Crowley J, Alison Arnold S, Wood N, et al. Monitoring a high cell density recombinant *Pichia pastoris* fed-batch bioprocess using transmission and reflectance near infrared spectroscopy [J]. Enzyme and Microbial Technology, 2005, 36(5 ~ 6): 621 ~ 628.

8 Tamburini E, Vaccari G, Tosi S, et al. Near-infrared spectroscopy: a tool for monitoring submerged fermentation processes using an immersion optical-fiber probe [J]. Applied Spectroscopy, 2003, 57(2): 132 ~ 138.

9 Sáiz-Abajo M J, González-Sáiz J M, Pizarro C. Prediction of organic acids and other quality parameters of wine vinegar by near-infrared spectroscopy. A feasibility study [J]. Food Chemistry, 2006, 99(3): 615 ~ 621.

10 Cozzolino D, Kwiatkowski M J, Parker M, et al. Prediction of phenolic compounds in red wine fermentations by visible and near infrared spectroscopy [J]. Analytica Chimica Acta, 2004, 513(1): 73 ~ 80.

11 Schölkopf B, Smola A J, Williamson R C, et al. New support vector algorithms [J]. Neural Computation, 2000, 12(5): 1 207 ~ 1 245.

12 张庆, 徐光华, 华成, 等. 一类支持向量机的设备状态自适应报警方法[J]. 西安交通大学学报, 2009, 43(11): 61 ~ 65.
Zhang Qing, Xu Guanghua, Hua Cheng, et al. Self-adaptive alarm method for equipment condition based on one-class support vector machine [J]. Journal of Xi'an Jiaotong University, 2009, 43(11): 61 ~ 65. (in Chinese)

13 Mourão-Miranda J, Hardoon D R, Hahn T, et al. Patient classification as an outlier detection problem: an application of the one-class support vector machine [J]. NeuroImage, 2011, 58(3): 793 ~ 804.

14 Wold S, Esbensen K, Geladi P. Principal component analysis [J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1 ~ 3): 37 ~ 52.

15 Thissen U, Pepers M, Üstün B, et al. Comparing support vector machines to PLS for spectral regression applications [J]. Chemometrics and Intelligent Laboratory Systems, 2004, 73(2): 169 ~ 179.