文章编号： 1672-8785(2010)01-0030-06

# 应用近红外光谱快速鉴别不同年龄段人食用的奶粉品种

唐玉莲 梁逸曾 范 伟 张华秀

(中南大学化学化工学院中草药现代化研究中心，湖南长沙 410083)

**摘 要**：应用近红外光谱分析技术 (NIRS) 并结合支持向量机 (SVM)，对三种不同年龄段人食用的奶粉品种进行了鉴别。先采用 Kennard-Stone 法对 150 个样本进行挑选，选出 120 个作为训练集，剩余的 30 个作为预测集。实验中选用径向基函数 (RBF) 为核函数，采用二次网格搜索和五折交叉验证优化两个建模参数：核参数 $\gamma$ 和惩罚因子 $C$，最佳值为 $\gamma$=0.03125，$C$=2048。用最优参数值建立的校正模型，对训练集和预测集的判别率均可达到 100%。与主成分分析 (PCA) 进行了比较。结果表明，SVM 鉴别准确率高于 PCA，说明近红外光谱可以快速、准确地鉴别不同年龄段人食用的奶粉品种。

**关键词**：近红外光谱；支持向量机；奶粉；鉴别

**中图分类号**： O657 **文献标识码**： A **DOI**: 10.3969/j.issn.1672-8785.2010.01.007

# Fast Discrimination of Varieties of Different Age Rank Milk Powder Using Near Infrared Spectroscopy

TANG Yu-Lian, LIANG Yi-Zeng, FAN Wei, ZHANG Hua-Xiu

*(Research Center of Modernization of Chinese Herbal Medicines, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China)*

**Abstract:** Three kinds of different age rank milk powder are identified by using the near infrared spectroscopy with support vector machines (SVM). First, the Kennard-Stone method is used to select 120 training sets from 150 samples and other 30 samples are used as the prediction sets. In the experiment, the radial basis function is selected as the kernel function and the two-step grid searching and five-fold cross validation are used to optimize two model parameters: kernel function $\gamma$ and penalty factor C. The optimal $\gamma$ and C are 0.03125 and 2048 respectively. The correction model established with the optimal parameters has an identification rate of 100% for both training sets and prediction sets. By comparison with the principal component analysis (PCA), the SVM exhibits its higher identification accuracy. This shows that the near infrared spectroscopy can identify the varieties of different age rank milk powder quickly and accurately.

**Key words:** NIRS; Support Vector Machine; milk powder; discrimination

## 1 引言

奶粉的主要成分有蛋白质、脂肪酸、碳水化合物、维生素、矿物质、多种免疫物质和牛磺酸等。目前，人们主要通过用化学方法测定奶粉的化学成分来鉴别奶粉的品质 [1−3]，这种方法的

**收稿日期**：2009–09–02

**基金项目**：国家自然科学基金 (20875104， 10771217) 资助项目；2008 年湖南省标准化战略资助项目

**作者简介**：唐玉莲 (1982-)，女，湖南邵阳人，在读研究生，主要研究方向为近红外快速检测奶粉及液态奶品质。 E-mail:tangyulian0824@163.com

检验结果准确，但过程耗时费力，不利于快速检测。因此，寻找一种快速有效的奶粉质量检测方法很有必要。

近红外光谱分析技术具有成本低、无污染、无破坏性、测试重现性好、样品无需预处理并可在线检测以及多组分同时测定等优点，已被越来越多地应用于食品工业、石油化工、农业等领域 [4] 。国内外很多学者对利用近红外光谱技术鉴别物质品种作过研究 [5−8] 。

支持向量机是一种新的机器学习方法 [9] ，它通过核函数实现高维空间的非线性映射，较好地解决了小样本、非线性和高维模式识别等实际问题，并有效克服了神经网络学习方法中网络结构难以确定、收敛速度慢、局部极小点、过学习与欠学习以及训练时需要大量数据样本等不足，具有良好的泛化能力，在很多领域得到了广泛应用。在化学领域中，支持向量机的应用不断增加，显示了其独特优势。文献 [10–12] 报道了支持向量机在食品质量控制中的应用， U.Thissen 等人研究了支持向量机在化学反应监控中的应用 [13] ，李洪东等人 [14] 综述了支持向量机的基本原理及其在化学中的应用。本文尝试把近红外光谱分析技术与支持向量机方法结合起来，对三种不同适龄奶粉品种进行鉴别，以期为奶粉质量控制提供新的方法。

## 2 实验部分

### 2.1 仪器设备

采用的仪器设备为 Antaris Ⅱ傅里叶变换近红外光谱分析仪 (由美国 Thermo Nicolet 公司生产) 、漫反射积分球附件和 InGaAs 检测器。

### 2.2 样品来源及光谱测量

婴儿、青少年、中老年三种不同适龄阶段的奶粉购自长沙沃尔玛、家乐福、大润发超市，包括伊利、蒙牛、雅士利、南山、雀巢和惠氏等多个奶粉厂家的不同批次产品，每种适龄阶段奶粉各收集 50 个样本，共 150 个样本。

把一定量的样品放在专用的测量杯中，以仪器内部的空气为背景， 用积分球和旋转台测定所有样品的 NIR 漫反射光谱。仪器参数设定如下：扫描范围为 $10000\text{cm}^{-1} \sim 4000\text{cm}^{-1}$ ，分辨率为 $8\text{cm}^{-1}$ ，扫描信号累加 64 次，每个样品平行测定 3 次，取平均光谱。图 1 是 10 个中老年奶粉、10 个青少年奶粉和 10 个婴儿奶粉品种的三次平均光谱图。

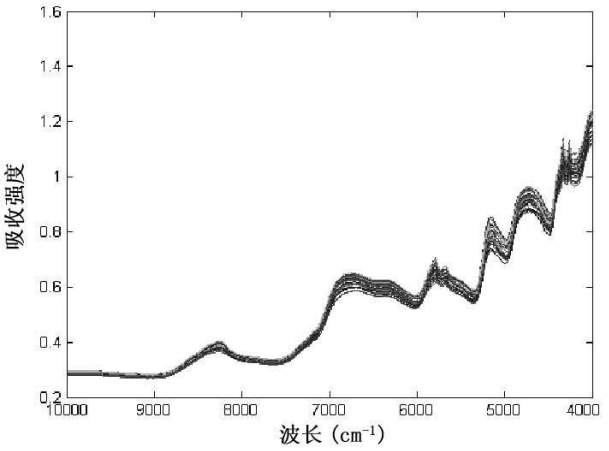



图 1　部分奶粉的 NIR 光谱图

Libsvm 算法共享程序包由台湾大学林智仁提供 (http://www.csie.ntu.edu.tw/cjlin/libsvm) 。其他程序采用 MATLAB 6.5 自行编写，在 Pentium Ⅳ 216GHz 处理器上运行。

## 3 基本原理

### 3.1 Libsvm 的基本原理

Libsvm 分类包括 C–SVC 和 nu–SVC 两种，本研究采用 C–SVC 。其原理如下：对于二元分类，如果训练向量使 $x_i \in R^n$ ， $i = 1, \cdots, l$ 分为两类，向量 $y \in R^l$ 有 $y_i \in \{1, -1\}$ ， C–SVC [15] 解决以下优化问题： $\frac{1}{2}\omega^T\omega + C\sum_{i=1}^{20}\xi_i$ ，约束条件为 $y_i[\omega^T\phi(x_i) + b] \geq 1 - \xi_i$ ， $\xi_i \geq 0$ ， $i = 1, \cdots, 20$ ，它的对偶问题为 $\frac{1}{2}\alpha^T Q\alpha - p^T\alpha$ ， $0 \leq \alpha_i \leq C$ ， $i = 1, \cdots, l$ ， $y^T = 0$ ，其中 $p$ 代表所有向量之一， $C > 0$ 是拉格朗日乘数的上界， $Q$ 是 $l \times l$ 阶的半正定矩阵，并且 $Q_{ij} \equiv y_i y_j K(x_i, x)$ ， $K(x_i, x) \equiv \phi(x_i)^T\phi(x_j)$ 是核函数。训练向量 $x_i$ 被函数 $\phi$ 映射到高维空间，如图 2 所示。决策函数为 $sign\left[\sum_{i=1}^{l} y_i\alpha_i K(x_i, x) + b\right]$ 。

对于多元分类， Libsvm 使用“一对一”的方法进行多元分类。应用这种方法分类需要构建 $k(k-1)/2$ 个二元分类器，每个分类器对两个不同的训练数据集进行分类。为了训练第 i 类和第 j 类数据集中的训练数据，需要解下面的二元分类问题：

$$\min_{\omega^{ij},b^{ij},\xi^{ij}} \frac{1}{2}(\omega^{ij})^T\omega^{ij} + C\left[\sum_t (\xi^{ij})_t\right]$$

$[(\omega^{ij})^T\phi(x_t)] + b^{ij} \geq 1 - \xi_t^{ij}$ ，如果 $x_t$ 在 $i$ 类，$[(\omega^{ij})^T\phi(x_t)] + b^{ij} \leq -1 + \xi_t^{ij}$ 。如果 $x_t$ 在 $j$ 类，则在分类过程中， Libsvm 使用投票策略：将每次二元分类看成是一次投票过程，将每个数据点看成是一个选票，最后该数据点被归类为得票最多的类。如果两类具有相同的票数，由于没有找到更好的方法， Libsvm 认为它属于索引序号较小的类。

支持向量机中使用的核函数主要有四类：线性核函数： $K(X_i, X_j) = X_i^T X_j$ ；多项式核函数： $K(X_i, X_j) = (\gamma X_i^T X_j + r)^d$ ， $\gamma > 0$ ；RBF 核函数： $K(X_i, X_j) = \exp(-\gamma \|X_i - X_j\|^2)$, $\gamma > 0$ ；Sigmoid 核函数： $K(X_i, X_j) = \tan h(\gamma X_i^T X_j + r)$ ，其中， $\gamma$ 、 $r$ 和 $d$ 均为核参数。

### 3.2 Kennard-Stone 的原理

Kennard-Stone 法是把所有的样本都看作训练集候选样本，从中依次挑选样本进入训练集[16] 。首先，选择欧氏距离最远的两个向量对进入训练集；定义 $d_{ij}$ 为从第 i 个样本向量到第 j 个样本向量的欧氏距离，假设已有 k 个样本向量被选进训练集，这里 k 小于样本总数 n ，针对第 v 个待选样本向量，定义最小距离

$$D_{kv} = \min(d_{1v}, d_{2v}, \cdots, d_{kv})$$

所有待选样本向量的 $D_{kv}$ 最大值 $D_{mkv} = \max(D_{kv})$ ，拥有最大最小距离 $D_{mkv}$ 的那条待选样本进入训练集。依此类推，达到要求的样本数目。该方法的优点是能保证训练库中的样本按照空间距离分布均匀，缺点是需要进行数据转换和计算量大。

## 4 结果与讨论

### 4.1 主成分分析 (PCA)

主成分分析 (PCA) 是模式识别判别分析中最常用的一种线性映射方法，该方法的中心目的是将数据降维，对原变量进行变换，使少数几个新变量成为原变量的线性组合。 PCA 结合不同光谱预处理获得的训练集的前两个主成分的得分分布，如图 3(a) 、 (b) 、 (c) ，分别为训练集原始光谱、经过多元散射 (MSC) 和一阶微分预处理后的 PCA 的得分分布图，其中 * 代表中老年奶粉样本中的 39 个训练集样本， ○代表青少年奶粉样本中的 40 个训练集样本， △代表婴儿奶粉样本中的 41 个训练集样本。

由图 3 可以看出， (a) 图中训练集样本的光谱未经过光谱预处理。在用 PCA 进行判别时，

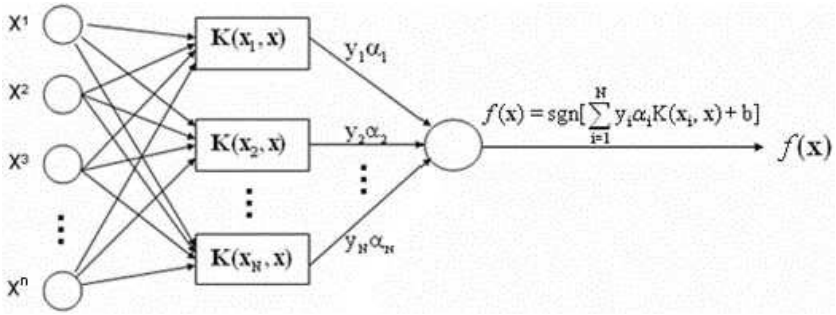
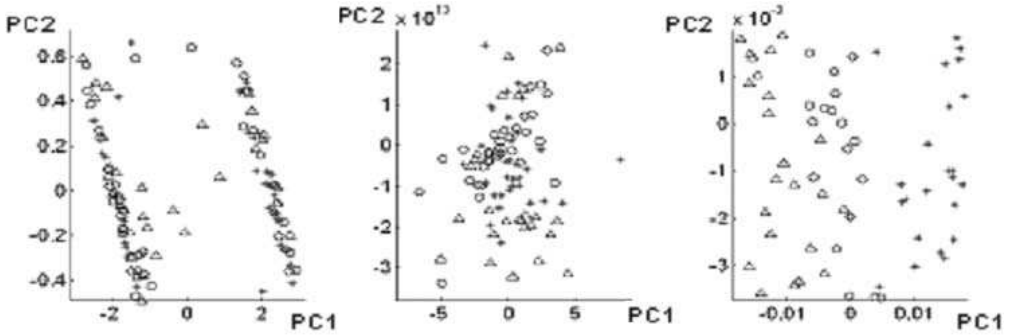



图 2 支持向量机的结构图 ($x = [x^1, x^2, x^3, \cdots, x^n]$ 表示输入向量)




(a)　　(b)　　(c)

图 3 (a) 训练集原始光谱 PCA 前两个主成分的得分分布；(b) 训练集光谱经过 MSC 预处理后 PCA 前两个主成分的得分分布；(c) 训练集光谱经过一阶微分预处理后 PCA 前两个主成分的得分分布

出现了严重的混淆，三类样本基本不能分开。(b) 图采用 MSC 对训练集样本进行预处理后，判别结果稍微有所改进，三类样本还是不能分开。(c) 图采用一阶微分处理后，三类样本有分开的趋势，青少年奶粉和中老年奶粉基本上分开了，但是婴幼儿奶粉和青少年奶粉仍然有一部分夹杂在一起，不能完全分别开来，而当用 SVM 对三类样本进行鉴别时，即使光谱不经过预处理，也能得到很好的分类结果。结果表明，本实验中样本品种之间存在内在差异，采用 PCA 方法虽然能够有效地去除噪声和降低特征维数，保留了描述样本分布的特征，但在降维的同时也丢失了很多有用信息，并没有保留最有利于分类的特征。SVM 通过核函数将输入空间 (低维) 中的线性不可分数据映射成高维特征空间线性可分数据，这样就可以实现线性分类，在保留样本最大特征信息的同时，对样本作了很好的分类，故本文采用支持向量机。

### 4.2 Kennard-Stone(KS) 法训练集和测试集的选取

选择有代表性的训练集不但可以减少建模的工作量，而且可以直接影响所建模型的适用性和准确性。本文共有 150 个样本，中老年奶粉依次编号为 (1–50)，青少年奶粉依次编号为 (51–100)，婴幼儿奶粉依次编号为 (101–150)，所有光谱数据都经过中心化处理。表 1 是通过 Kennard-Stone 法依次挑选出来的 30 个测试集样本，依次编号为 1–30(由于训练集样本数多，这里只列出预测集)。

表 1 利用 KS 法选出的 30 个预测集样本的序号

| 样品编号 | 样品索引 |
|---|---|
| 1–10 | 6 8 9 11 12 13 14 16 17 35 |
| 11–20 | 46 54 59 62 67 69 71 72 74 75 |
| 21–30 | 83 110 111 114 115 118 128 132 133 135 |

### 4.3 参数选择

采用支持向量机时需要确定的参数有两个，一是合适的核函数，二是最佳核函数参数。而现今还没有系统的方法论可用来选择核函数。通过与其他核函数的比较，得知高斯径向基函数 (RBF) 作为非线性函数能够减少训练过程中计算的复杂性，因此本文选用 RBF 为核函数。

在采用 RBF 核函数的 SVM 模型中，惩罚因子 $C$ 和核参数 $\gamma$ 的选择是非常重要的。$C$ 用于衡量误差罚分的大小，对改进 SVM 模型非常重要。$\gamma$ 则控制函数回归误差，并且会直接影响初始的特征值和特征向量。$\gamma$ 过小，会导致过拟合。相反，$\gamma$ 过大，会导致欠学习。此外，$\gamma$ 还关系到 SVM 模型对输入变量噪声的灵敏度。本研究中采用二步格点搜索法和交叉验证相结合的方法来选择参数。具体做法如下：(1) 设定

参数 $C$ 和 $\gamma$ 的起始值、步长、终止值。这里，$C$ 按起始值、步长、终止值依次设为 [5,2,15]，$\gamma$ 设为 [5,–2,–15]；(2) 运用指数序列搜索方法 ($C = 2^{-5}, 2^{-3}, \cdots, 2^{15}$；$\gamma = 2^{5}, 2^{3}, \cdots, 2^{-15}$) 逐格搜索每一组 $(C,\gamma)$ 格点的值，根据每个格点得到的交叉验证准确率来确定最优取值。在搜索过程中，用五折交叉验证的最高准确率可达 99.1667%，故本文采用五折交叉验证。通过对 $C$ 和 $\gamma$ 逐格搜索，训练集在 $C$ =2048 和 $\gamma$ =0.03125 即 $\log_2 C$ =11 、 $\log_2 \gamma = -5$ 时的预测准确率最优，为 99.1667%。图 4 显示了对训练集进行参数优化的结果，表 2 列出了寻优过程中部分参数对 $(C,\gamma)$ 的值及对应的交叉验证准确率。

### 4.4 模型建立及预测

利用最优参数 c=2048、g=0.03125 建立 SVM 校正模型，对训练集的预测准确率为 100%，对 30 个预测集的预测准确率为 100%，预测结果见表 3。

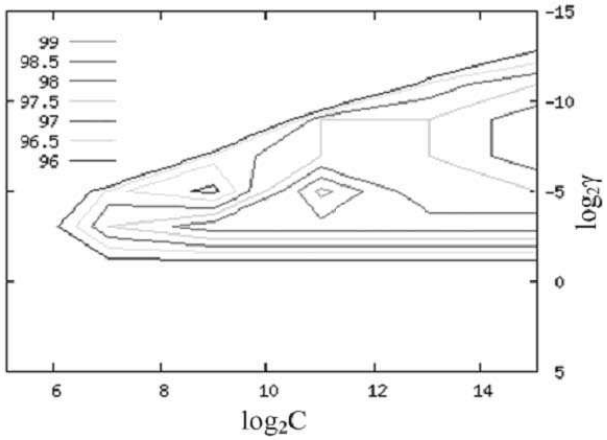



图 4 训练集的参数 $C$ 和 $\gamma$ 的优化结果图

表 2 7 组参数对 $(C,\gamma)$ 的值及对应的交叉验证准确率

| $C$ | $\gamma$ | 交叉验证准确率 |
|---|---|---|
| $2^3$ | $2^{-1}$ | 93.3333% |
| $2^5$ | $2^{-3}$ | 94.1667% |
| $2^3$ | $2^1$ | 95.8333% |
| $2^7$ | $2^{-5}$ | 96.6667% |
| $2^7$ | $2^{-3}$ | 97.5000% |
| $2^9$ | $2^{-3}$ | 98.3333% |
| $2^{11}$ | $2^{-5}$ | 99.1667% |

表 3 30 个预测集的预测结果

| 样品编号 | 真实值 | 预测值 | 样品编号 | 真实值 | 预测值 | 样品编号 | 真实值 | 预测 |
|---|---|---|---|---|---|---|---|---|
| 6 | 1 | 1 | 8 | 1 | 1 | 9 | 1 | 1 |
| 11 | 1 | 1 | 12 | 1 | 1 | 13 | 1 | 1 |
| 14 | 1 | 1 | 16 | 1 | 1 | 17 | 1 | 1 |
| 35 | 1 | 1 | 46 | 1 | 1 | 54 | 2 | 2 |
| 59 | 2 | 2 | 62 | 2 | 2 | 67 | 2 | 2 |
| 69 | 2 | 2 | 71 | 2 | 2 | 72 | 2 | 2 |
| 74 | 2 | 2 | 75 | 2 | 2 | 83 | 2 | 2 |
| 110 | 3 | 3 | 111 | 3 | 3 | 114 | 3 | 3 |
| 115 | 3 | 3 | 118 | 3 | 3 | 128 | 3 | 3 |
| 132 | 3 | 3 | 133 | 3 | 3 | 135 | 3 | 3 |

## 5 结论

本研究采用支持向量机与近红外光谱相结合的方法，选用径向基核函数，利用二步网格搜索法优化参数，对 3 种不同适龄奶粉品种的 150 个样本进行了鉴别，取得了满意的分类结果。该方法具有简单、速度快、准确、无损等优点，可为奶粉生产过程的质量控制提供一种新的方法。

## 参考文献

[1] 吴迪，何勇，冯水娟，等. 基于 LS–SVM 的红外光谱技术在奶粉脂肪含量无损检测中的应用 [J]. **红外与毫米波学报**, 2008, **27**(3): 180–184.

[2] Wu Di, He Yong, Feng Shuijuan. Short-wave near-infrared spectroscopy analysis of major compounds in milk powder and wavelength assignment [J]. *Analytica Chimica Acta*, 2008, **610**: 232–242.

[3] Borin A, Ferrão M F, Mello C, et al. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk [J]. *Analytica Chimica Acta*, 2006, **579**: 25-32.

[4] 陆婉珍，袁洪福，徐广通，等. **现代近红外光谱分析技术（第二版）**[M]. 北京：中国石化出版社，2006.

[5] Esteban-Diez I, Gonzalez-Saiz J M, Pizarro C. An evaluation of orthogonal signal correction methods for the characterisation of arabica and robusta coffee varieties by NIRS [J]. *Analytica Chimica Acta*, 2004, **514**(1): 57–67.

[6] Zsolt Seregely, Tamas Deak, Gyorgy Denes Bisztray. Distinguishing melon genotypes using NIR spectroscopy [J]. *Chemometrics and Intelligent Laboratory Systems*, 2004, **72**(2): 195–203.

[7] 柴金朝，金尚忠. 近红外光谱技术在纺织布料聚类分析中的应用 [J]. **红外**, 2009, **30**(1): 31–35.

[8] LIU Fei, HE Yong. Classification of brands of instant noodles using Vis/NIR spectroscopy and Chemometrics [J]. *Food Research International*, 2008, **41**: 562–567.

[9] Vapnik V. *Statistical learning theory* [M]. New York: Wiley, 1998.

[10] A Borin, M F Ferrao, C Mello, et al. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk [J]. *Anal. Chim. Acta*, 2006, **579**(1): 25–32.

[11] J A Fernandez Pierna, P Volery, R Besson, et al. Classification of modified starches by Fourier transform infrared spectroscopy using support vector machines [J]. *Agric. Food. Chem*, 2005, **53**(17): 6581–6585.

[12] K Brudzewski, S Osowski, T Markiewicz. Classification of milk by means of an electronic nose and SVM neural network [J]. *Sensors and Actuators B*, 2004, **98**: 291–298.

[13] U Thissen, M Pepers, B U stun, et al. Comparing support vector machines to PLS for spectral regression applications [J]. *Chemometr. Intell. Lab. Syst*, 2004, **73**(2): 169–179.

[14] LI Hong-dong, LIANG Yi-zeng, XU Qing-song. Support vector machines and its applications in chemistry [J]. *Chemometrics and Intelligent Laboratory Systems*, 2009, **95**: 188–198.

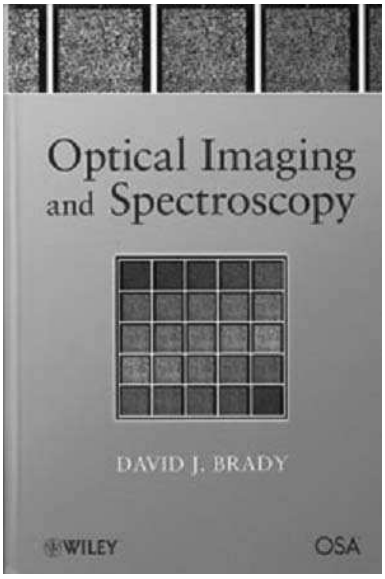
[15] Hsu C W, Chang C C, Lin C J. A Practical Guide to Support Vector Classification. http:// www.csie.ntu.edu.tw/ ~ cjlin/papers/guide/guide.pdf.

[16] Kennard RW, Stone LA. Computer aided design of experiments [J]. *Technometrics*, 1969, **11**: 137–148.

**新闻动态 News**

## 约翰威立公司出版传感器系统设计新书

据《Photonics Spectral》杂志报道，约翰威立出版有限公司联合美国光学学会于 2009 年出版了一本书名为《光学成像与光谱学》的光学传感器系统设计参考书。该书共 528 页，售价为 110 美元，内容涵盖以下几个方面：编码孔径与层析成像；成像系统设计中的相干测量策略；光场的几何模型、波模型和统计模型；现代光学探测器和焦平面阵列的基本功能；以及光学系统中的采样和变换，包括数字系统分析必需的小波和广义采样技术。




□ 高国龙