文章编号 1004-924X(2014)09-2352-07

用二维相关近红外谱和多维主成分分析判别掺杂牛奶

杨仁杰[1*],刘 蓉[2],杨延荣[1],张伟玉[1]

(1. 天津农学院 工程技术系,天津 300384;
2. 天津大学 精密测试技术及仪器国家重点实验室,天津 300072)

摘要:为了有效地提取牛奶中微量的掺杂物特征信息,提出了基于二维相关近红外光谱多维主成分分析(MPCA)和最小二乘支持向量机(LS-SVM)判别牛奶掺杂物的方法。首先,采集纯牛奶、掺杂尿素牛奶和掺杂三聚氰胺牛奶的一维近红外谱,并对其进行相关计算,构建各样品的二维相关近红外谱。然后,采用多维主成分分析法分析二维相关谱矩阵,压缩数据,提取相关谱的得分矩阵。最后,将提取的得分矩阵输入最小二乘支持向量机,分别建立掺杂尿素牛奶、掺杂三聚氰胺牛奶及两种掺杂牛奶与纯牛奶的 LS-SVM 判别模型。用所建模型对测试集未知样品进行了判别,结果显示其判别正确率分别为 92.3%,96.2%,92.3%。研究结果表明:所提出的方法不仅有效提取了牛奶中掺杂物的特征信息,而且缩短了建模所需时间,取得了较好的判别效果。

关 键 词:二维相关近红外光谱;多维主成分分析;掺杂牛奶;尿素;三聚氰胺

中图分类号:O657.33 **文献标识码**:A **doi**:10.3788/OPE.20142209.2352

Classification of adulterated milk by two-dimensional correlation near-infrared spectroscopy and multi-way principal component analysis

YANG Ren-jie[1*], LIU Rong[2], YANG Yan-rong[1], ZHANG Wei-yu[1]

(1. *College of Engineering and Technology, Tianjin Agricultural University, Tianjin* 300384, *China*;
2. *State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin* 300072, *China*)
* *Corresponding author, E-mail: rjyang*1978@163.*com*

Abstract: To extract effectively characteristic information of adulterants in milk, the classification models for adulterated milk were established using two-dimensional (2D) correlation near-infrared spectra combining a Multi-way Principal Component Analysis (MPCA) with Least Square Support Vector Machines (LS-SVM). First, one-dimensional near-infrared spectra of pure milk and adulterated milk samples were collected and the synchronous 2D correlation spectra of all samples were calculated. Then, the MPCA was used to reduce dimension by extracting score matrix of 2D correlation data set.

收稿日期:2013-11-05;**修订日期**:2013-12-31.

基金项目:国家自然科学基金资助项目(No. 31201359, No. 81471698);天津市自然科学基金资助项目(No. 13JCYBJC25700)

Finally, LS-SVM models for urea-tainted milk, melamine-tainted milk, and the above two kinds of adulterated milk were constructed by using score matrix extracted from 2D correlation spectra as the input variables. These models were used to discriminate the known samples in the test set and the results show that the classification accuracy rates of unknown samples are 92.3%, 96.2%, 92.3%, respectively. It demonstrates that the proposed method not only extracts effectively feature information of adulterants in milk, but also reduces the input dimension of LS-SVM and computational time. It realizes a better classification of adulterated milk and pure milk.

Key words: two-dimensional correlation near-infrared spectroscopy; multi-way principal component analysis; adulterated milk; urea; melamine

1　引　言

近年来，食品安全问题屡见不鲜，从“瘦肉精事件”到“苏丹红咸鸭蛋”，从“镉米”到“地沟油”，从“2008 年三聚氰胺奶粉”到“2013 年新西兰奶粉”，这些事件暴露出我国食品安全问题的严重性。目前，食品安全问题已成为社会关注的焦点，研究便捷快速的检测方法也成为维护食品安全的重要课题[1]。

近红外光谱作为一种快速、无损的检测技术，已经被广泛用于乳制品、蜂蜜、食用油、酒类等食品的掺杂检测中。由于食品中掺杂物的多样化和微量化，加之掺杂物的特征峰与食品固有组分特征峰相互重叠，因此一维近红外光谱无法有效地解析图谱。二维相关谱相对一维谱提高光谱分辨率，可有效地解析图谱[2-3]；同时，二维相关谱体现的是随外扰变化的特征信息，因此可以根据研究目的选择特定的外扰构建二维相关谱，提高光谱的选择性[3]。目前，二维相关谱已广泛应用于各种复杂生物体系的组分分析。

清华大学孙素琴等利用二维相关光谱技术获得高分辨的“指纹图谱”，并提出了比较完整的红外宏观指纹图谱法和红外三级鉴定方法[4-6]。文献[7-9]直接将二维相关谱技术结合多维化学计量学来实现掺杂牛奶的定性、定量分析，并与一维光谱的预测结果进行比较，指出基于二维相关谱的预测结果优于一维谱。但二维相关谱是一个三维矩阵，包含着大量的数据信息，有些数据对分类的贡献很小或因测量原因而有可能不稳定。若将全部数据都用于建模，不仅会影响模型的准确性，还增加了运算量、降低了运算速率。文献[10-11]基于统计方法参数化理论提取了二维相关同步谱统计参数，并将所提取的参数与模式识别方法相结合建立了掺杂牛奶判别模型，并取得了较好的判别效果。

主成分分析是特征提取与数据降维的常用方法。多维主成分分析（Multi-way Principal Component Analysis，MPCA）是二维主成分分析方法应用于多维数据阵的扩展[12-13]。本文提出了一种基于二维相关近红外谱，并结合 MPCA 和最小二乘支持向量机（Least Square Support Vector Machines，LS-SVM）的掺杂牛奶判别方法。该方法既可有效提取二维相关谱的特征信息，又可减少建模时间。

2　材料与方法

2.1　实验仪器与样品处理

实验所用仪器为美国 PerkinElmer 公司的傅里叶变换红外光谱仪。该仪器采用卤素灯光源，石英（Quartz）分束器和液氮冷却的锑化铟（InSb）检测器，控制光信号衰减的 B-Stop 和 J-Stop 分别为 1.5，3.96 mm。其光谱采集范围为 4 000～10 000 cm^{-1}，分辨率为 4 cm^{-1}，扫描 8 次求平均值。

实验中纯牛奶样本选购自本地某超市，包括蒙牛纯牛奶、伊利纯牛奶和海河纯牛奶。随机选取上述纯牛奶，分别配置不同质量浓度的掺杂尿素牛奶（1～20 g/L）和掺杂三聚氰胺牛奶（0.01～3 g/L）样品各 40 个，纯牛奶样品 80 个。在采集样品光谱数据前，需对掺杂牛奶样品进行匀质处理。蒸馏水的光谱作为背景光谱，采用样品光谱扣除相邻背景光谱后的数据来计算二维相关谱。

2.2　二维相关近红外同步谱

假设原始一维近红外光谱矩阵 $\mathbf{S}(k\times m)$ 中

包含 k 个光谱，根据二维相关 Noda 理论[2]，则同步二维相关谱可表示为[7-11, 14]：

$$\boldsymbol{\Phi}(\nu_1,\nu_2)=\frac{1}{k-1}\mathbf{S}^{\mathrm{T}}\mathbf{S}. \tag{1}$$

本文实验中 k 取 2，即光谱矩阵 $\mathbf{S}$ 中包括 2 个光谱：第一行为纯牛奶一维近红外光谱；当第二行为第 i 个掺杂牛奶或纯牛奶的一维近红外光谱时，根据式(1)就可得到第 i 个掺杂牛奶或纯牛奶所对应的二维相关同步谱。

2.3 多维主成分分析法

实验时，从原始一维谱计算二维相关谱矩阵 $\mathbf{X}(I\times J\times K)$，其中 I 为样品数，J 和 K 为波数变量。用 MPCA 算法首先将二维相关谱矩阵 $\mathbf{X}$ 沿着波数轴进行切分(见图 1)，构成一个新的矩阵 $\mathbf{X}(I\times JK)$，然后同二维主成分分析一样，用 MPCA 将 $\mathbf{X}$ 分解为得分向量 $\boldsymbol{t}_r$ 与载荷向量 $\boldsymbol{p}_r$ 的乘积[15]，并加上残差矩阵 $\boldsymbol{E}$，从而有：

$$\mathbf{X}=\sum_{r=1}^{R}\boldsymbol{t}_r\otimes\boldsymbol{p}_r^{\mathrm{T}}+\boldsymbol{E}, \tag{2}$$

式中：R 为主成分个数，⊗为 Kronecker 乘积。

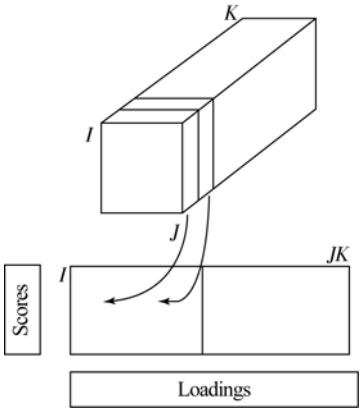

图 1　MPCA 计算过程

Fig. 1　Computation procedure of MPCA method

2.4 最小二乘支持向量机

LS-SVM 是经典支持向量机的一种改进，与其它线性或非线性的支持向量机算法相比，LS-SVM降低了计算复杂性，提高了回归精度[10,16]。因此，LS-SVM 在许多分类问题中都有很好的应用效果。LS-SVM 的具体算法可参考文献[16]和[17]。

在 LS-SVM 模型中，核函数和模型参数的选择至关重要。本文采用非线性的径向基函数(Radial Basis Function，RBF)为核函数，采用网格搜索方法，以交叉验证均方根误差(Root Mean Square Error of Cross-validation，RMSECV)为最小化目标函数，优选两个重要的模型参数 γ(正则化参数)和 σ^2(核函数 RBF 参数)。

2.5 数据处理

采用自行编写的二维相关分析的 Matlab 程序对所有样品进行二维相关计算，得到各样品所对应的二维相关同步谱。由于二维相关谱的矩阵数据较大，采用 MPCA 对其进行降维，将所提取的主成分输入 LS-SVM(LS-SVM 算法工具包参见 Suykens 等提供的网络共享 http:// www.esat.kuleuven.ac.be/sista/lssvmlab)建立掺杂牛奶与纯牛奶的判别模型。所有的计算都采用 MATLAB2012a 软件工具(Mathwork Inc.)来完成。

3 结果与分析

文献[11]研究了纯牛奶、掺杂尿素牛奶及掺杂三聚氰胺牛奶在 4 400～4 800 $\mathrm{cm^{-1}}$ 的二维近红外相关谱特性，并指出虽然二维相关谱相对于一维谱可提高光谱的分辨率，可揭示被覆盖或被淹没的弱峰。但由于掺杂牛奶是复杂的生物体系，纯牛奶与掺杂牛奶的二维相关近红外谱也非常相似，只有在细微处存在差别，因此无法通过肉眼正确识别掺杂牛奶与纯牛奶。

将 40 个掺杂尿素牛奶的近红外相关谱矩阵(40×51×51)、40 个掺杂三聚氰胺牛奶的近红外相关谱矩阵(40×51×51)和 80 个纯牛奶的近红外相关谱矩阵(80×51×51)合并为一个新的矩阵 $\mathbf{X}$(160×51×51)。采用 MPCA 方法分析 $\mathbf{X}$，图 2 是两种掺杂牛奶与纯牛奶样品在前 3 个主成分空间散点的分布图。显然，纯牛奶类样品、掺杂尿素牛奶类样品，掺杂三聚氰胺牛奶类样品按其种类分布在不同的区域，但不同种类样品之间的界限比较模糊，且部分样品在类别边界处存在重叠，无法直接从图中加以区别。为了更准确地识别掺杂牛奶和纯牛奶，需要借助化学计量学手段。

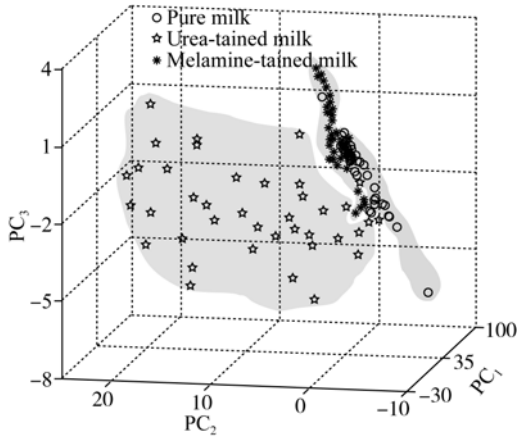

图 2　160 个样品在前三主成分空间散点分布

Fig. 2　3-D score plots of MPCA results for 160 samples by PC_1, PC_2 and PC_3

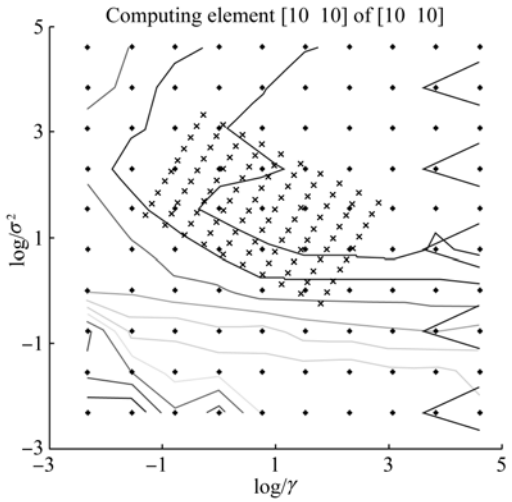

图 3　LS-SVM 参数寻优等高线图

Fig. 3　Contour plot of optimization error for LS-SVM

在对 160 个样品二维相关红外谱矩阵进行 MPCA 的基础上，将前 4 个主成分得分矩阵（前 4 个主成分的累计贡献率为 99%，说明前 4 个主成分代表原始二维相关谱矩阵的大多数信息）输入 LS-SVM，建立两种掺杂牛奶与纯牛奶的判别模型。

采用浓度梯度法从 80 个掺杂牛奶（掺杂尿素牛奶、掺杂三聚氰胺牛奶各 40 个）和 80 个纯牛奶样品中选出 108 个（掺杂尿素牛奶和掺杂三聚氰胺牛奶各 27 个，纯牛奶 54 个）作为训练集，余下 52 个样品作为独立的测试集。在训练集和测试集中，纯牛奶和掺杂牛奶分别用“－1”，“1”来表示其类别属性。

在建立 LS-SVM 模型的过程中，模型参数 γ 和 σ^2 的选择是至关重要的一步。为了确定最优的建模参数 γ 和 σ^2，分别计算在不同参数 γ 和 σ^2 组合下 LS-SVM 模型的 RMSECV。采用二步网格搜索和交叉验证相结合对两个模型参数组合进行全局寻优。模型参数 γ 和 σ^2 的初始值都为 0.1，搜索范围都为 0.1～100。首先采用较大步长网格搜索整个区域（用“·”表示），得到全局最优区域；然后缩小搜索区域为全局最优区域，以较小步长搜索（用“×”表示），搜索过程见图 3。图中所示曲线为不同参数组合 log2(γ) 和 log2(σ^2) 下，LS-SVM 模型的 RMSECV 等高线。当 RMSECV 最小时，所对应的 γ 和 σ^2 即为最优的建模参数组合。根据图 3，最终确定最优的模型参数：γ=2.7 和 σ^2=4.7。

LS-SVM 的参数确定后，通过 LS-SVM 建立掺杂牛奶与纯牛奶的判别模型。利用所建立的模型对训练集中的样品进行内部预测，预测结果见图 4。54 个纯牛奶样品都得到正确识别，判别正确率为 100%；各有 2 个掺杂尿素牛奶和掺杂三聚氰胺牛奶被误判，两种掺杂牛奶的判别正确率都为 92.6%。在训练集 108 个样品中，仅有 4 个样品发生误判，因此所建模型对训练集样品内部预测的判别正确率为 96.3%。

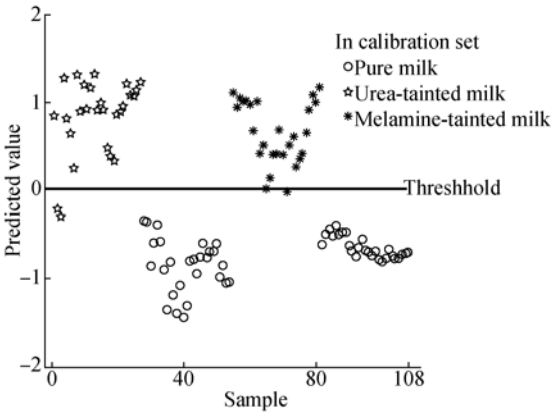

图 4　LS-SVM 模型对训练集样品的预测结果

Fig. 4　Values of samples in train set predicted by LS-SVM model

利用所建立的 LS-SVM 模型对测试集中的未知样品进行外部预测，其预测结果见图 5。在测试集中，共有 2 个样品被误判（掺杂尿素牛奶和掺杂三聚氰胺牛奶各 1 个），26 个纯牛奶都得到

了正确识别，因此所建模型对测试集样品的总的判别正确率为 92.3%。

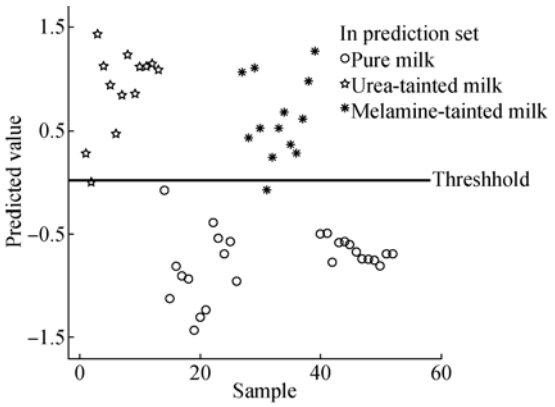

图 5　LS-SVM 模型对测试集样品的预测结果

Fig. 5　Values of samples in test set predicted by LS-SVM model

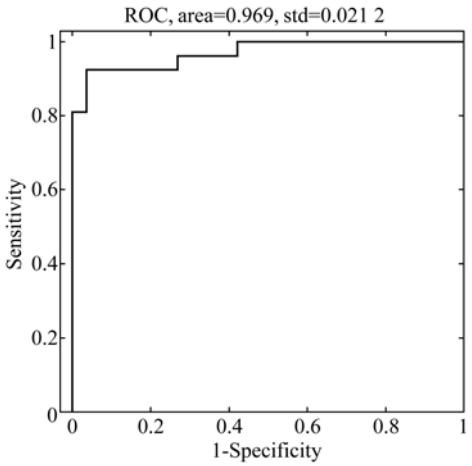

图 6　掺杂牛奶与纯牛奶分类的 ROC 曲线

Fig. 6　Curve of receiver operating characteristic (ROC) of adulterated milk and pure milk

为了进一步评价基于 MPCA 分析所建立的 LS-SVM 模型对测试集样品的判别效果，对其进行了(Receiver Operating Characteristic，ROC)分析。ROC 分析是将灵敏度和特异度结合起来综合评价判别效果的一种方法。ROC 曲线下的面积(Area Under Curve，AUC)越接近于 1，说明模型的判别效果越好，当 AUC 介于 0.7～0.9 时，说明模型具有一定的判别效果；当 AUC 大于 0.9 时，说明模型具有较高的判别效果[16,18]。图 6 是两种掺杂牛奶与纯牛奶判别的 ROC 曲线，AUC=0.969，表明通过 MPCA 提取相关谱矩阵后建立的 LS-SVM 模型具有较强的预测效果。

同时，也对两种掺杂牛奶与各自对应纯牛奶的二维相关近红外谱进行了 MPCA 分析，提取了前 4 个主成分(前 4 个主成分的累计贡献率都超过 99%)，将提取的主成分输入 LS-SVM 分别建立掺杂尿素牛奶和掺杂三聚氰胺与纯牛奶的判别模型。在两个模型中，训练集都由 54 个样品(纯牛奶和掺杂牛奶各 27 个)组成，测试集都由 26 个样品(纯牛奶和掺杂牛奶各 13 个)组成。表 1 给出了两个 LS-SVM 的建模参数和判别结果。对于掺杂尿素牛奶的 LS-SVM 模型，所有纯牛奶样品都得到了正确识别，在训练集和测试集中各有 2 个掺杂牛奶被误判，模型对训练集和测试集的判别正确率分别为 96.3%，92.3%。对于掺杂三聚氰胺牛奶的 LS-SVM 模型，在训练集和测试集中各有 1 个掺杂牛奶被误判，所建模型对训练集和测试集样品的判别正确率分别为 98.1%，96.2%。

表 1　两个 LS-SVM 模型建模参数及预测结果

Tab. 1　Parameters and discrimination results of two LS-SVM models

模型	建模参数	判别正确率	
		训练集	测试集
掺杂尿素牛奶	$\gamma=23.35$ $\sigma^2=12.25$	96.3%	92.3%
掺杂三聚氰胺牛奶	$\gamma=0.52$ $\sigma^2=14$	98.1%	96.2%

4　结　论

本文将二维相关近红外光谱技术与 MPCA 结合起来，首先测得纯牛奶和掺杂牛奶的常规一维近红外谱，并对各样品进行相关计算得到各样品对应的二维相关近红外谱。接着采用 MPCA 提取并压缩了牛奶中掺杂物的特征信息，并将所提取的信息输入 LS-SVM，分别建立了掺杂尿素牛奶、掺杂三聚氰胺牛奶及两种掺杂牛奶与纯牛奶的判别模型。所建模型对测试集未知样品的判别正确率分别为 92.3%，96.2%，92.3%。该方法不仅提取了待分析物质的特征信息，而且也解决了二维相关谱数据量大，建模时间长的问题。本课题组也在开展用该方法检测未知掺杂物牛奶的相关研究。

参考文献:

[1] 杨延荣,杨仁杰,张志勇,等. 红外光谱结合核隐变量正交投影法判别掺杂牛奶[J]. 光学 精密工程, 2013, 21(10):77-84.
YANG Y R, YANG R J, ZHANG ZH Y, *et al.*. Discrimination of adulterated milk based on infrared spectroscopy and K-OPLS [J]. *Opt. Precision Eng.*, 2013, 21(10):77-84. (in Chinese)

[2] NODA I. Advances in two-dimensional correlation spectroscopy [J]. *Vibrational Spectroscopy*, 2004, 36(2):143-165.

[3] 杨仁杰. 基于二维相关谱掺杂牛奶检测方法研究[D]. 天津:天津大学,2013.
YANG R J. *Research on the detection of adulterated milk based on 2D correlation spectroscopy* [D]. Tianjin: Tianjin University, 2013. (in Chinese)

[4] SUN S Q, ZHOU Q, CHEN J B. *Infrared Spectroscopy for Complex Mixtures: Applications in Food and Traditional Chinese Medicine* [M]. Beijing: Chemical Industry Press, 2011.

[5] WANG Y, XU C H, WANG P, *et al.*. Analysis and identification of different animal horns by a three-stage infrared spectroscopy [J]. *Spectrochim Acta Part A: Molecular and Biomolecular Spectroscopy*, 2011, 83(1):265-270.

[6] ZHANG Y L, CHEN J B, LEI Y, *et al.*. Discrimination of different red wine by Fourier-transform infrared and two-dimensional infrared correlation spectroscopy [J]. *Journal of Molecular Structure*, 2010, 974(1-3):144-150.

[7] YANG R J, LIU R, XU K X. Detection of adulterated milk using two-dimensional correlation spectroscopy combined with multi-way partial least squares[J]. *Food Bioscience*, 2013, 2:61-67.

[8] 杨仁杰, 刘蓉, 徐可欣, 等. 二维相关近红外谱结合 NPLS-DA 判别掺杂牛奶的研究[J]. 光子学报, 2013, 42(5):580-585.
YANG R J, LIU R, XU K X, *et al.*. Discrimination of adulterated milk using NPLS-DA combined with two-dimensional correlation near-infrared spectroscopy [J]. *Acta Photonica Sinica*, 2013, 42(5):580-585. (in Chinese)

[9] YANG R J, LIU R, XU K X, *et al.*. Discrimination of adulterated milk based on two-dimensional correlation spectroscopy (2D-COS) combined with kernel orthogonal projection to latent structure (K-OPLS) [J]. *Applied Spectroscopy*. 67(12):1363-1367.

[10] YANG R J, LIU R, XU K X, *et al.*. Classification of adulterated milk with the parameterization of 2D correlation spectroscopy and least squares support vector machines [J]. *Anal. Methods*, 2013, 5: 5949-5953.

[11] 苗静,曹玉珍,杨仁杰,等. 基于二维相关近红外谱参数化及 BP 神经网络的掺杂牛奶鉴别[J]. 光谱学与光谱分析, 2013, 33(11):3032-3035.
MIAO J, CAO Y ZH, YANG R J, *et al.*. Identification of adulterated milk based on two-dimensional correlation near-infrared spectra parameterization and BP neural network [J]. *Spectroscopy and Spectral Analysis*, 2013, 33(11):3032-3035. (in Chinese)

[12] GELADI P, ISAKSSON H, LINDQVIST L, *et al.*. Principal component analysis of multivariate image [J]. *Chemometrics and Intelligent Laboratory Systems*, 1989, 5(3):209-220.

[13] SMILDE A K, DOORNBOS D A. Three-way methods for the calibration of chromatographic systems: comparing PARAFAC and three-way PLS[J]. *Journal of Chemometrics*, 1991, 5: 345-360.

[14] CHEN J B, ZHOU Q, NODA I, *et al.*. Quantitative classification of two-dimensional correlation spectra [J]. *Applied Spectroscopy*, 2009, 63(8): 920-925.

[15] VILLEZ K, RUIZ M, SIN G, *et al.*. Combining multiway principal component analysis (MPCA) and clustering for efficient data mining of historical data sets of SBR process [J]. *Water Science and Technology*, 2008, 57(10):1659-1666.

[16] 虞科, 程翼宇. 一种基于最小二乘支持向量机算法的近红外光谱判别分析方法[J]. 分析化学,

2006, 34(4):561-564.

YU K, CHENG Y Y. Discriminating the genuineness of Chinese medicines with least squares support machines [J]. *Chinese Journal of Anaytica Chemistry*, 2006, 34(4):561-564. (in Chinese)

[17] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers [J]. *Neural Process. Lett.*, 1999, 9(3):293-300.

[18] LASKO T A, BHAGWAT J G, ZOU K H. The use of receiver operating characteristic curves in biomedical informatics [J]. *Journal of Biomedical Informatics*, 2005, 38(5):404-415.

作者简介：

杨仁杰(1978－)，男，山西运城人，博士，讲师，2005 年于南开大学获得硕士学位，2013 年于天津大学获得博士学位，主要从事食品安全检测方面的研究工作。E-mail:rjyang1978@163.com

(版权所有　未经许可　不得转载)