基于近红外透射光谱分析和BP神经网络的大豆品种识别

杨冬风[1],朱洪德[2]

(1. 黑龙江八一农垦大学 信息技术学院,黑龙江 大庆 163319;2. 黑龙江八一农垦大学 农学院,黑龙江 大庆 163319)

摘 要:为了实现大豆品种的快速无损鉴别,对16份大豆品种的近红外透射光谱(NITS)进行分析。首先通过平滑和马氏距离的光谱预处理方法消除噪声和去除奇异光谱。然后分别用主成分分析(PCA)和离散多带小波变换(DWT)提取光谱特征,作为BP神经网络的输入,构建PCA-BP和DWT-BP大豆品种识别模型。结果表明:PCA-BP模型的识别准确率为98.125%,平均识别时间为9.3 ms;DWT-BP模型的识别准确率为95.93%,平均识别时间为6.4 ms。研究结果为大豆品种的快速无损鉴别提供了理论依据和实用方法。

关键词:近红外透射光谱;主成分分析;离散小波变换;BP神经网络;大豆

中图分类号:TP391.4 **文献标识码**:A **文章编号**:1000-9841(2013)02-0249-05

Recognition of Soybean Varieties Based on Near Infrared Transmittance Spectroscopy and BP Neural Network

YANG Dong-feng[1], ZHU Hong-de[2]

(1. Department of Information Technology, Heilongjiang Bayi Agricultural University, Daqing 163319, China; 2. Agronomy College, Heilongjiang Bayi Agricultural University, Daqing 163319, China)

Abstract: In order to realize rapid nondestructive recognition of soybean varieties, near infrared transmittance spectrum (NITS) of 16 soybean samples were analyzed. Smoothing treatment and Mahalanobis distance were used to filter noise and wipe off singular spectrum. Principal component analysis (PCA) and discrete wavelet transform (DWT) were respectively used to extract spectral features which act as the input of BP neural network. PCA-BP and DWT-BP identification model were built. The accuracy rate of PCA-BP model and DWT-BP model were 98.125% and 95.93%, in addition, the average recognition time were 9.3 ms and 6.4 ms. The results of the investigation provided the theoretical support and practical method for rapid nondestructive recognition of soybean varieties.

Key words: Near infrared transmittance spectroscopy; Principal component analysis; Discrete wavelet transform; BP Neural network; Soybean

大豆是重要的油料和高蛋白粮饲兼用作物。含有丰富的蛋白质、脂肪和多种人体有益的生理活性物质。黑龙江省大豆种植历史悠久,具有地理、生态和经济优势,大豆常年种植面积约366.4万 hm^2,占全国大豆种植面积的1/3,是我国大豆的主产区。1996~2011年,通过黑龙江省农作物品种审定委员会审定的大豆品种达132个,这些新品种的推广应用对黑龙江省乃至东北地区大豆生产的发展起到了重要推动作用。随着大豆品种数量的增加,种子市场鱼龙混杂,给国家和农民利益造成很大损失。快速、科学、准确地鉴别大豆品种,对保护农民利益,保障国家粮食安全以及种子质量鉴定、品种权益保护均具有重要意义。

近红外光谱(NIRS)是利用有机质在近红外光谱区的振动吸收从而快速测定样品中多种化学成分含量的一项技术。NIRS的主要信息来源是含氢基团O-H、N-H、C-H的倍频和组合频,光谱信息丰富,适合多组分测定。品种的光谱鉴别依据的是不同品种在光谱特征空间分布的特征,国内外学者在近红外品种识别领域开展了诸多研究[1-4]。

大豆品种识别相关研究目前并不多见。洪庆红等[5]利用傅里叶变换红外光谱(FTIR)法测定了8个大豆品种的子叶及外表皮的红外光谱,认为不同大豆品种的FTIR在1 800~1 200 cm^{-1}范围内有较大的差异,可以用于鉴别不同的大豆品种,但并未给出具体的鉴别方案,是探索性的研究。朱大洲等[6]根据单粒大豆的近红外光谱,结合软独立建模分类法(SIMCA)建立定性分析模型,可以准确鉴别垦鉴豆43和中黄13,虽然鉴别的准确率高,但识别范围较窄。

本研究使用近红外透射光谱技术对大豆群体品种样本进行分析,采用主成分分析和离散小波变

收稿日期:2012-12-10
基金项目:引进国际先进农业科学技术计划"948计划"(2008-Z24)。
第一作者简介:杨冬风(1977-),女,硕士,讲师,主要从事模式识别研究。E-mail:yangsansun@sina.com。
通讯作者:朱洪德(1962-),男,博士,研究员,主要从事大豆遗传育种和栽培研究。E-mail:Zhd495@sohu.com。

换两种光谱数据特征提取与 BP 神经网络相结合的方法建立大豆品种的近红外光谱识别模型,并对两种模型的识别结果进行对比和分析。

1 试验材料

供试材料为绥农 4 号、绥农 10 号、绥农 21、合丰 25、合丰 38、垦农 16、垦鉴 43、九丰 6 号、东农 42、抗线虫 4 号、宝丰 8 号、北丰 9 号、北丰 14、北丰 15、黑农 44 和黑农 41 共 16 份大豆材料,由黑龙江省农业科学院大豆研究所提供。

2011 年采集,成熟收获后风干保存。试验前,样品统一筛选,每个品种称取 2 000 g 备用。每个大豆品种均分成 40 份,其中 20 份作为训练集,剩余 20 份作为验证集。以品种分组,总训练集 320 份,总验证集 320 份。

2 识别模型的构建

近红外技术分为漫反射技术和透射技术。漫反射技术典型的谱区应用范围是 1 100 ~ 2 500 nm,这段谱区的近红外光在样品中的穿透一般不大于 1 mm,所以非常适合均质良好的粉状样品分析。透射技术典型的谱区应用范围是 700 ~ 1 100 nm,为分子振动吸收的二级和三级倍频吸收区。透射谱区的近红外光在样品的穿透最大可达 30 mm,因此适合样品的整粒或原状分析,是完全的无损分析技术[7-8]。为了实现快速、无损的大豆品种识别,本研究采用近红外光谱透射技术对不同品种的大豆样本进行分析。

2.1 光谱采集

采用 FOSS Infratec 1241 型谷物品质分析仪采集光谱数据,扫描波长范围为 850 ~ 1 050 nm,扫描步长为 0.4 nm,每个光谱的数据点 500 个,扫描次数 64 次。为消除样品粒度大小、均匀性不一致等因素对光谱的影响,每小份样品重复装样 3 次,取平均光谱作为该样品的光谱数据。16 个品种大豆的典型近红外透射光谱曲线如图 1 所示。

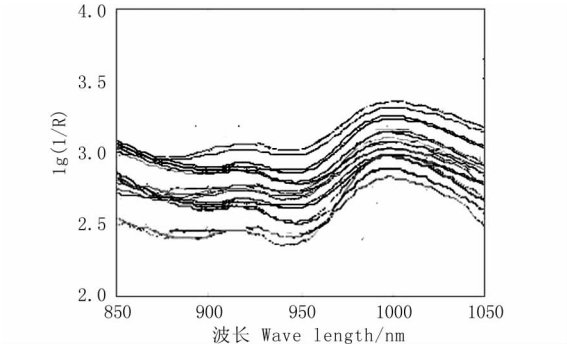

图 1 16 个品种大豆的近红外透射光谱谱线

Fig. 1 Near infrared transmittance spectrum of 16 soybean varieties

由于外界电磁干扰或采集情况的细微变化都会引起光谱数据的噪声或出现奇异光谱,因此,需要对原始光谱数据进行滤波去噪和筛选。本研究采用移动窗口平滑去噪方法和马氏距离奇异光谱筛选方法对原始光谱数据进行预处理。光谱数据预处理以及特征提取和模型的建立均采用 Matlab 7.0 实现。

2.2 特征提取

光谱数据预处理后,提取能表征品种类别的光谱空间分布特征作为 BP 神经网络的输入。研究了两种特征提取方法,一种是主成分分析特征提取,把采用这种提取方式建立的模型称为 PCA-BP 模型;另一种特征提取方法是离散多带小波变换,把采用这种特征提取方式建立的模型称为 DWT-BP 模型。

2.2.1 *主成分分析(PCA)* 主成分分析的目的是将数据降维,以消除众多信息共存中相互重叠的信息部分。方法是通过对实测的多个指标相关矩阵内部结构关系的研究,构造少数几个综合的主成分指标,它们都是原指标的线性组合,不仅保留原指标的主要信息,且互不相关,同时比原指标具有更好的性质。

试验采集的大豆图谱共有 500 个数据点,数据量大,冗余信息多。利用 Matlab 中的主成分分析函数 princomp( )对数据矩阵 $X_{320\times500}$ 进行主成分分析,分析得出各主成分的特征值、每个主成分对方差的贡献率和累积贡献率。从累积贡献率可以看到,前 16 个主成分已经提取了 99.211% 的方差(表 1),说明这 16 个变量可以很好地表征原数据的信息。因此,对每个训练集样本的光谱数据提取 16 个主成分进行回归模型的建立。

表 1 前 16 个主成分及其贡献率

Table 1 The first 16 principal components and contribution rate

主成分 Principal component	贡献率 Contribution/%	主成分 Principal component	贡献率 Contribution/%
PC1	32.348	PC9	0.733
PC2	20.612	PC10	0.468
PC3	15.123	PC11	0.387
PC4	13.025	PC12	0.354
PC5	5.314	PC13	0.325
PC6	4.258	PC14	0.306
PC7	3.406	PC15	0.294
PC8	2.005	PC16	0.253

2.2.2　离散多带小波变换(DWT)特征提取　每一条光谱所包含的品种信息,都体现在谱线的波动以及波峰波谷的细微变化中。原则上构成整个波形的 500 个数据点都可以作为 BP 神经网络识别模型的输入,然而众多的特征值会给识别模型造成巨大的计算量。由于小波变换在时域和频域同时具有良好局部化性质,把光谱数据作为离散信号,用离散多带小波变换对光谱进行特征提取,可以有效地实现数据降维,同时保留各光谱数据点所包含的品种信息。

M 带离散小波变换,包括 1 个低通滤波器和(M-1)个高通滤波器,滤波系数分别记为:

$$l=(l_0,l_1,\cdots,l_{N_f-1}) \quad h^{(s)}=(h_0^{(s)},h_1^{(s)},\cdots,h_{N_f-1}^{(s)}) \tag{1}$$

滤波器从 l 级到 l+1 级离散小波变换系数有:

$$c_{l+1,k}=\sum\frac{N_f-1}{i=0}l_i c_{l,mk+i} \quad d_{l+1,k}^{(s)}=\sum\frac{N_f-1}{i=0}h_i^{(s)}c_{l,mk+i} \tag{2}$$

其中,$c_{l,k}$表示尺度系数,$d_{l,k}^{(s)}$表示小波系数,N_f为滤波系数长度。

从 M 带离散小波变换的一级到下一级,只有低通滤波器有输出,各带中下一级的系数数目是上一级的 1/M。滤波器把输入数据分解成不同频带,用(l,t)带表示第 l 级上第 t 带($t\in0,1,\cdots,m$-1)的离散小波变换,用 $x^{[1]}(t)$表示(l,t)带的离散小波变换系数。

对于每份样本的光谱数据,数据点的数目为 500,每 2 点取一点作为离散小波变换的采样信号($c_{0,0},c_{0,1},c_{0,2},\cdots,c_{0,250}$)。采用 5 带小波变换,$N_f=5$。250 个信号($c_{0,0},\cdots,c_{0,250}$)的 5 带小波变换如图 2 所示。此小波变换含有 1 个低通滤波器(L)和 4 个高通滤波器(H_1,H_2,H_3,H_4)。

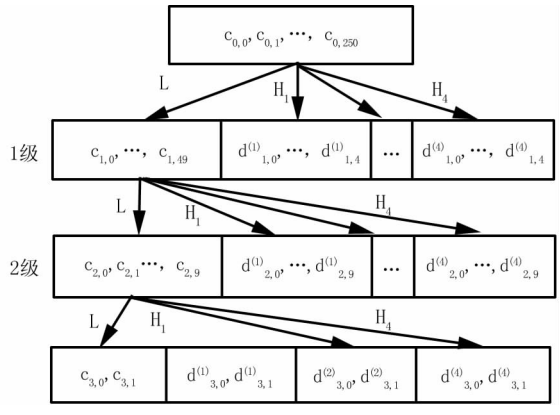

图 2　离散小波变换

Fig. 2　A 5-band discrete wavelet transform for 250 sample signals

取(3,0)带上的离散小波变换系数 $c_{3,0},c_{3,1},d_{3,0}^{(1)},d_{3,1}^{(1)},d_{3,0}^{(2)},d_{3,1}^{(2)},d_{3,0}^{(3)},d_{3,1}^{(3)}$共 8 个特征参数作为神经网络的输入。

2.3　BP 神经网络识别模型构建

2.3.1　输入、输出节点的确定　BP 神经网络是一种按误差逆传播算法训练的多层前馈网络,它的学习规则是使用最速下降法,通过反向传播来不断调整网络的权值和阈值,使网络的误差平方和最小。BP 神经网络模型拓扑结构包括输入层、输出层和 1 个或多个隐含层。

输入层起缓冲存储器的作用,把数据源加到网络上。其节点数目取决于输入特征向量的维数。采用主成分分析方法共提取了 16 个特征,把它们归一化后作为表征大豆品种的向量输入到神经网络中,因此输入层共有 16 个节点。采用离散多带小波变换对光谱数据降维,共提取了 8 个特征,同样对它们进行归一化处理后作为识别模型的输入,因此输入层共有 8 个节点。

输出层节点数一般为要识别的种类数,也可以用输出节点的编码表示。本研究中输出种类 16 个,对应 16 个大豆品种,把输出设计为 4 位二进制数(表 2)。

表 2　大豆类别对应的二进制输出值

Table 2　The binary output value corresponding to the soybean variety

品种名称 Variety	输出 Output	品种名称 Variety	输出 Output
绥农 4 号 Suinong 4	0000	东农 42 Dongnong 42	1000
绥农 10 号 Suinong 10	0001	抗线虫 4 号 Kangxianchong 4	1001
绥农 21 Suinong 21	0010	宝丰 8 号 Baofeng 8	1010
合丰 25 Hefeng 25	0011	北丰 9 号 Beifeng 9	1011
合丰 38 Hefeng 38	0100	北丰 14 Beifeng 14	1100
垦农 16 Kennong 16	0101	北丰 15 Beifeng 15	1101
垦鉴 43 Kenjian 43	0110	黑农 44 Heinong 44	1110
九丰 6 号 Jiufeng 6	0111	黑农 41 Heinong 41	1111

2.3.2　隐含层节点的确定　BP 网络中,隐含层节点数的选择对网络的性能影响很大。隐含层节点过少,则局部极小值多,不能达到网络训练的目的;隐含层节点数过多,网络拟合函数过于复杂,容易产生过拟合,使得网络的泛化能力变差,这样即使网络训练效果很好,验证效果也可能很差。本研究中隐含层节点的确定方法是先用估算公式计算出初始节点数,然后根据训练情况动态调整隐含层节点数目到最佳数目为止[9]。

节点估算公式如公式(3):

$$h=\sqrt{0.43mn+0.12n^2+2.54m+0.77n+0.35}+0.51 \tag{3}$$

其中,h 为估算的隐含层节点数,m 为输入节点

数，n 为输出节点数。

采用主成分分析提取特征时，估算出的初始隐含节点数为 9 个。经过一定次数的训练调整，最终确定较合适的隐含层节点数为 10 个。

采用离散多带小波变换提取特征时，估算初始隐含节点数为 7 个。经过训练调整，最终确定较合适的隐含层节点数为 9 个。

2.3.3 *学习算法的确定* S 型函数是激活函数 φ(·)的最佳选择，如 logisic 函数、双曲正切函数等。S 型函数一般具有光滑、可微、非线性和饱和等特性，而且导函数 φ′(·)容易用 φ(·)本身来表达，计算简单。本研究中，隐含层神经元的激活函数选用 logsic( ) 函数，输出层的激活函数选 pureline( ) 函数。学习规则采用 Delta 学习规则，即按误差函数梯度下降的方向进行权值的调整。

2.4 BP 神经网络训练过程及结果讨论

640 组光谱数据预处理后，每份样本分别按 PCA 及 DWT 提取出 16 个和 8 个特征并做归一化运算后备用。将备用数据分为两大组，其中 320 组作为训练集，剩余的 320 组作为验证集。用训练集中的数据对构建的神经网络进行训练，经过多次试验，系统设置训练总步长为 960，每隔 8 步显示 1 次，PCA-BP 模型的均方误差(MSE)为 0.122 6，模型的平均识别时间为 9.2 ms，DWT-BP 模型的 MSE 为 0.372 8，模型的平均识别时间为 6.4 ms。将 320 组验证数据输入训练好的网络，将得到的二进制值与已知类别编码对照，计算这 320 组数据的识别准确率。PCA-BP 模型的识别率为 98.125%，DWT-BP 模型的识别率为 95.93%(表 3)。

表 3 模型识别结果

Table 3 Recognition results

验证分组 Test group	PCA-BP			DWT-BP		
	样本数目 Sample size	正确识别 Correct recognition	识别率 Recognition rate/%	样本数目 Sample size	正确识别 Correct recognition	识别率 Recognition rate/%
绥农 4 号 Suinong 4	20	20	100	20	19	95
绥农 10 号 Suinong 10	20	19	95	20	19	95
绥农 21 Suinong 21	20	20	100	20	20	100
合丰 25 Hefeng 25	20	19	95	20	18	90
合丰 38 Hefeng 38	20	20	100	20	20	100
垦农 16 Kennong 16	20	20	100	20	19	95
垦鉴 43 Kenjian 43	20	20	100	20	20	100
九丰 6 号 Jiufeng 6	20	20	100	20	19	95
东农 42 Dongnong 42	20	20	100	20	20	100
抗线虫 4 号 Kangxianchong 4	20	20	100	20	20	100
宝丰 8 号 Baofeng 8	20	20	100	20	19	95
北丰 9 号 Beifeng 9	20	20	100	20	19	95
北丰 14 Beifeng 14	20	18	90	20	19	95
北丰 15 Beifeng 15	20	19	95	20	18	90
黑农 44 Heinong 44	20	20	100	20	19	95
黑农 41 Heinong 41	20	19	95	20	19	95

从验证试验结果看，PCA-BP 模型有 6 组误判，误判的品种有绥农 10 号、合丰 25、北丰 14、北丰 15 以及黑农 41，其中绥农 10 号、合丰 25、北丰 15 和黑农 41 各误判 1 组，北丰 14 误判 2 组。DWT-BP 模型有 13 组误判，绥农 4 号、绥农 10 号、垦农 16、九丰 6 号、宝丰 8 号、北丰 9 号、北丰 14、黑农 44 和黑农 41 各误判 1 组，合丰 25 和北丰 15 各误判 2 组。

3 讨 论

采用近红外透射光谱测量了 16 个大豆品种的光谱数据，利用主成分分析和离散小波变换两种方法提取光谱特征，分别建立了 BP 神经网络品种识别模型。结果表明，两种模型的识别精度都达到了大豆品种初筛的要求，PCA-BP 模型的识别效果优

于 DWT-BP 模型,但 DWT-BP 模型的平均识别时间比 PCA-BP 模型短 2.8 ms。主成分分析在提取特征时充分考虑了数据的内部结构关系,更能体现出光谱数据在类别上的差异性;而离散小波变换在提取特征时,对数据的覆盖比较全面,但对数据间的关系和数据对体现类别的重要性上考虑较少,这是两种模型识别效果差异的根本原因。对于识别时间的长短,在神经网络层数一样的情况下,主要与各层的节点数目多少相关,PCA-BP 模型的输入节点和隐含节点均多于 DWT-BP 模型,因此其识别时间要长一些。

本研究建立的模型对于供试的大豆品种可以实现准确的识别,对于其他品种的大豆,需要对模型作适当的改变并进行有效的训练才能使用。

参考文献

[1] 周健,成浩,曾建明,等. 基于近红外的多相偏最小二乘模型组合分析实现茶叶原料品种鉴定与溯源的研究[J]. 光谱学与光谱分析,2010,30(10):2650-2653. (Zhou J,Cheng H,Zeng J M,et al. Combined analysis of multi-partial least squares models based on near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis,2010,30(10):2650-2653.)

[2] 曹芳,吴迪,何勇. 基于可见-近红外反射光谱技术的葡萄品种鉴别方法的研究[J]. 光学学报,2009,29(2):537-540. (Cao F,Wu D,He Y,et al. Variety disecrimination of grapes based on visible-near reflection infrared spectroscopy[J]. Acta Optica Sinica,2009,29(2):537-540.)

[3] 李晓丽,胡兴越,何勇. 基于主成分和多类判别分析的可见-红外光谱水蜜桃品种鉴别新方法[J]. 红外与毫米波学报,2006,25(6):417-420. (Li X L,Hu X Y,He Y. New approach of discrimination of varieties of juicy peach by near infrared spectra based on PCA and MDA model[J]. Journal of Infrared and Millimeter Waves,2006,25(6):417-420.)

[4] 陈兰珍,孙谦,叶志华,等. 基于神经网络的近红外光谱鉴别蜂蜜品种研究[J]. 食品科技,2009,34(8):287-290. (Chen L Z,Sun Q,Ye Z H,et al. Determination of floral origin of honey by near infrared spectroscopy based on artificial neutral network[J]. Food Science and Technology,2009,34(8):287-290.)

[5] 洪庆红,李丹婷,郝朝运. 应用 FTIR 直接测定法鉴定大豆的品种[J]. 光谱学与光谱分析,2005,25(8):1246-1249. (Hong Q H,Li D T,Hao C Y. Identification of soybean varieties by direct determination of FTIR spectrum[J]. Spectroscopy and Spectral Analysis,2005,25(8):1246-1249.)

[6] 朱大洲,王坤,周光华,等. 单粒大豆的近红外光谱特征及品种鉴别研究[J]. 光谱学与光谱分析,2010,30(12):3217-3221. (Zhu D Z,Wang K,Zhou G H,et al. The NIR Spectra based variety discrimination for single soybean seed[J]. Spectroscopy and Spectral Analysis,2010,30(12):3217-3221.)

[7] Villareal C P,Normita M,Cruz D L,et al. Rice amylose analysis by near-infrared transmittance spectroscopy[J]. Cereal Chem.,1992,71(3):292-296.

[8] Buchmannn N B,Josefsson H,Cowe I A. Performance of European artificial neural network(ANN) calibrations for moisture and protein in cereals using the danish near-infrared transmission(NIT) network[J]. Cereal Chemistry,1999,78(5):572-577.

[9] 王立琦. BP 神经网络在大豆油酸价近红外光谱检测中的应用[J]. 食品科学,2009,30(4):243-246. (Wang L Q. Application of BP neural network in detecting acid value of oil using near infrared spectrum[J]. Food Science,2009,30(4):243-246.)

(上接第 248 页)

参考文献

[1] Nakamura A,Furuta H,Maeda H,et al. Analysis of structural components and molecular construction of soybean soluble polysaccharides by stepwise enzymatic degradation[J]. Bioscience Biotechnology and Biochemistry,2001,65:2249-2258.

[2] Nakamura A,Furuta H,Maeda H,et al. The structure of soluble soybean polysaccharide[M]. Hydrocolloids,235-238.

[3] 前田裕一. 水溶性大豆多糖类の开发[J]. 食品与开发,1992,27(9):47-49. (Maeda Y. Development of soybean soluble polysaccharide[J]. Food Development,1992,27(9):47-49.)

[4] Matsumura Y,Li J. Effects of polysaccharides containing galacturonic acids on the dispersion stability of soy proteins[J]. Soy Protein Research,2006,9:53-57.

[5] 白卫东,王琴. 豆奶稳定性的研究[J]. 现代食品科技,2006,22(1):5-7. (Bai W D,Wang Q. Study on the stability of soybeanmilk[J]. Modern Food Science and Technology,2006,22(1):5-7.)

[6] Nakamura A,Yoshida R,Maeda H,et al. The stabilizing behaviour of soybean soluble polysaccharide and pectin in acidified milk beverage[J]. International Dairy Journal,2006,16:361-369.

[7] Liu J R,Nakamura A,Corredig M. Addition of pectin and soy soluble polysaccharide affects the particle size distribution of casein suspensions prepared from acidified skim milk[J]. Agricultural and Food Chemistry,2006,54:6241-6246.

[8] 黄来发. 蛋白饮料加工工艺与配方[M]. 北京:中国轻工业出版社,1996:86-87. (Huang L F. Protein beverage processing technology and formula[M]. Beijing: China Light Industry Press,1996:86-87.)

[9] 纪铁鹏,池永红. 均质对苦杏仁蛋白质溶解性的影响[J]. 科技与经济,2006(17):84-85. (Ji T P,Chi Y H. Effect of homogenization on the bitter almond protein solubility[J]. Science Technology and Economy,2006(17):84-85.)