

# 近红外光谱测定茶叶中茶多酚和茶多糖的人工神经网络模型研究

罗一帆<sup>1,2</sup>, 郭振飞<sup>3</sup>, 朱振宇<sup>4</sup>, 王川丕<sup>1</sup>, 江和源<sup>1</sup>, 韩宝瑜<sup>1\*</sup>

1. 农业部茶叶化学工程重点开放实验室, 浙江 杭州 310008
2. 华南师范大学化学与环境学院, 广东 广州 510631
3. 华南农业大学生命科学院, 广东 广州 510624
4. 中山大学基础医学院, 广东 广州 510089

**摘要** 为了建立近红外光谱测定茶叶中茶多酚和茶多糖的模型, 应用了人工神经网络方法, 选择了  $7\ 432.3\sim 6\ 155.7\ \text{cm}^{-1}$  和  $5\ 484.6\sim 4\ 192.5\ \text{cm}^{-1}$  特征光谱范围, 以网络结构参数的输入层、隐层、输出层神经元数目分别为(8, 4, 1)和(7, 5, 1)来建立茶多酚和茶多糖的测定模型, 模型的结果表明建模的茶多酚和茶多糖的  $r$ ,  $RMSECV$ ,  $RSECV$  分别为 0.984 7, 0.460, 0.123 和 0.947 0, 0.136, 0.224; 预测集的  $r$ ,  $RMSEP$ ,  $RSEP$  则分别为 0.980 4, 0.529, 0.017 和 0.968 2, 0.111, 0.030。由此说明建立的近红外光谱-人工神经网络模型可用于预测茶叶中茶多酚和茶多糖的含量。

**主题词** 人工神经网络; 近红外光谱; 茶多酚; 茶多糖

**中图分类号:** O657.1 **文献标识码:** A **文章编号:** 1000-0593(2005)08-1230-04

## 引言

近红外光谱介于可见光区和中红外光区之间, 波长范围为  $0.75\sim 2.5\ \mu\text{m}$ , 波数范围为  $4\ 000\sim 13\ 330\ \text{cm}^{-1}$ 。其光谱信息来源于有机分子中的含氢基团(C—H, O—H, N—H)振动的合频与各级倍频的吸收, 是样品中有机分子含氢基团的特征振动信息, 化学信息量相当丰富<sup>[1]</sup>。茶叶中的大多数有机成份如氨基酸、蛋白质、咖啡碱、茶多酚、还原糖、多糖、纤维素、半纤维素、淀粉、果胶等都含有各种氢基团, 所以通过对茶叶的近红外光谱分析可以测定这些成份的含量。然而我们要精确茶叶近红外谱带的归属很困难, 每个近红外谱带可能是若干个不同基频的倍频和合频谱带的组合, 没有锐峰和基线分离的谱峰, 大量的重叠谱峰和肩峰。而且这些光谱还包含键强度、化学组分、电负性和氢键信息<sup>[2, 3]</sup>。因此, 近红外分析是较复杂的弱信息分析, 这些信息多而烦琐, 如果选得不当, 就会丢掉许多有用光谱信息或出现共线问题, 建立的模型则不准确, 因此将化学计量学方法应用于近红外光谱分析之中, 有效而全面地分析光谱信息才能更好地应用近红外光谱数据, 人工神经网络方法就是其中一种<sup>[4]</sup>。

本文应用人工神经网络方法, 建立了茶叶中茶多酚和茶

多糖与近红外光谱之间的人工神经网络模型, 用内部交叉验证方法, 验证相关系数  $r$ 、标准偏差  $RMSECV$ 、预测标准偏差  $RMSEP$  和相对偏差  $RSEP$  等多个指标, 说明得到的预测效果是较好的。

## 1 人工神经网络原理与方法

人工神经网络方法是一种具有很强函数逼近能力的非线性建模方法, 已被应用于许多领域的研究中。为了得到更为精确的预测模型, 本文应用目前使用最广的误差反向传播神经网络(BP网络)算法<sup>[4, 5]</sup>进行计算, 采用了收敛速度快和学习时间短的 Levenberg-Marquardt 优化方法。计算时人工神经网络为 3 层网络结构, 首先选取近红外光谱数据, 全波段  $4\ 000\sim 9\ 000\ \text{cm}^{-1}$  总共有 1 297 个点。然后将光谱数据进行主成分样品变换, 取前几个主成分作为神经网络的输入, 分别建立近红外光谱与其茶多酚和茶多糖含量之间的相关模型, 网络输入层与隐含层、隐含层与输出层节点的传递函数用 Logsig 函数<sup>[6]</sup>。

为避免神经网络模型的过拟合, 本文使用交叉验证方法将样品分成训练集和预测集。随机抽取 20 个茶叶样本为预测样本集, 剩余 50 个样本为训练样本集。网络在训练集上进行建模, 在预测集上评价所建模型的准确性, 用相关系数  $r$

收稿日期: 2004-11-18, 修订日期: 2005-04-03

基金项目: 国家重点基础研究“973”计划(G1998051201), 广东省科技计划(2003C20405), 农业部茶叶化学工程重点开放实验室开放课题(200401)资助项目

作者简介: 罗一帆, 1962年生, 华南师范大学化学与环境学院副教授, 博士研究生 \* 通讯联系人

评价其相关性, 误差均方根  $RMSECV$  作为网络训练的评估标准, 预测误差均方根  $RMSEP$  反映模型对未知样本的预测效果。相关计算公式如下:

$$r = 1 - \frac{\sum (c_i - \hat{c}_i)^2}{\sum (c_i - c_m)^2} \quad (1)$$

$$RMSECV = \sqrt{\frac{\sum (\hat{c}_i - c_i)^2}{n - p}} \quad (2)$$

$$RMSEP = \sqrt{\frac{\sum (\hat{c}_i - c_i)^2}{m}} \quad (3)$$

其中  $c_i$  是实验值,  $\hat{c}_i$  是预测值,  $n$  是训练集样品数,  $p$  是使用的主成分数,  $m$  是预测集样品数。采用 (4) 式计算  $RSECV$  或  $RSEP$ , 以考察模型校正和预测的相对偏差。

$$RSECV, RSEP(\%) = \sqrt{\frac{\sum (\hat{c}_i - c_i)^2}{\sum c_i^2}} \times 100 \quad (4)$$

## 2 实验

### 2.1 仪器设备与试剂

Bruker IFS 28/N 傅里叶变换近红外光谱仪(德国布鲁克光学仪器公司), 积分球漫反射检测器, 环境温度为室温, 扫描次数为 64 次, 光谱分辨率为  $8 \text{ cm}^{-1}$ , 扫描范围为  $9\ 000 \sim 4\ 000 \text{ cm}^{-1}$ , OPUS 光谱采集软件。光谱数据以文本格式导出进行处理。神经网络等程序用 Matlab5.3 编制。

茶叶样品均由广东省农业科学院茶叶研究所提供。茶多酚和茶多糖含量测定所用的试剂均为分析纯。

### 2.2 茶叶中茶多酚和茶多糖含量的测定

茶叶中茶多酚和茶多糖含量均按国家标准(GB/T8313—1987)<sup>[7]</sup>进行测定。茶多酚采用酒石酸铁比色法, 茶多糖则采用蒽酮比色法。为了保证所测数据的准确性, 所有样品的测定均重复 6 次。测定结果表明所选择的茶叶样品含茶多酚为 22%~36%, 而茶多糖为 2.2%~4.5%。

### 2.3 近红外光谱数据的采集

样品为过 40 目筛的粉体, 置于光谱仪的积分球样品杯里, 扫描样品的近红外光谱。实验测定 70 个茶叶样品的近红外光谱, 如图 1 所示。

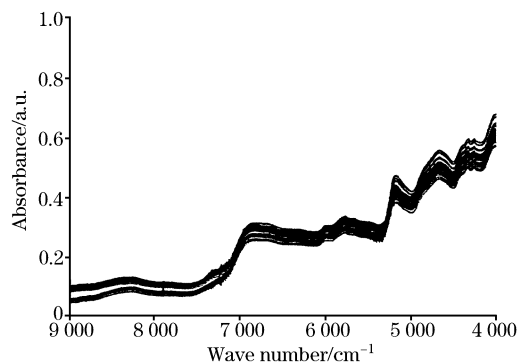


Fig. 1 NIR spectra of different tea samples

## 3 结果与讨论

### 3.1 神经网络参数的选择

神经网络参数主要有输入层、隐层、输出层神经元数目, 尤其是对输入层、隐层的节点数的选择尤为重要, 其数目太少会使网络模型不能收敛, 对未知样本的预测准确性差, 选择的数目太多又会产生过拟合现象<sup>[6]</sup>。为简化网络结构, 将近红外光谱矩阵的主成分作为神经网络输入, 并通过优化选择确定合适的输入层节点数(即光谱主成分数)和隐含层节点数<sup>[5]</sup>。输出层节点数定为 1, 即茶多酚和茶多糖含量值。为了得到最佳输入层、隐层的节点数, 我们对茶多酚和茶多糖近红外光谱模型在输入层节点数( $r$ )分别为 5~9 时, 不同隐层节点数(1~8)与其  $RMSECV$  值作图(见图 2), 由图 2(a)可知茶多酚的近红外光谱模型的输入层为 8, 隐层为 4 时, 其  $RMSECV$  值最低; 图 2(b)还显示茶多糖的近红外光谱模型的输入层为 7, 隐层为 5 时, 其  $RMSECV$  值最低。因此, 茶多酚和茶多糖近红外光谱模型的最佳网络结构其输入层、隐层、输出层神经元数目分别为(8, 4, 1)和(7, 5, 1)。

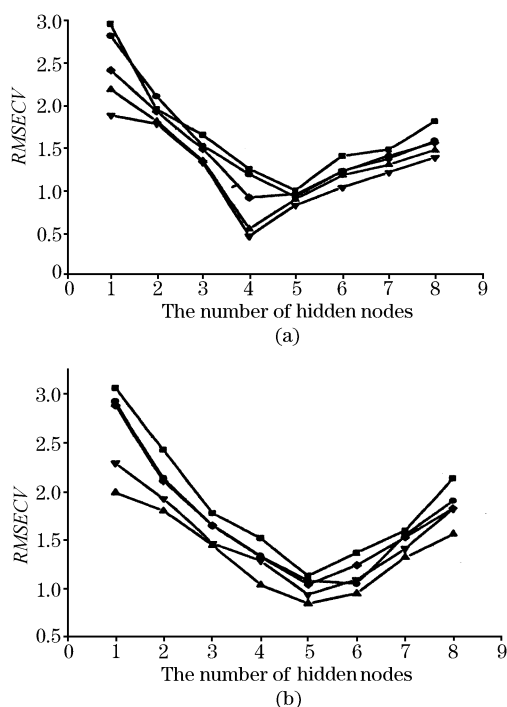


Fig. 2 Relationships between  $RMSECV$  values and the number of hidden nodes for (a) tea polyphenol and (b) tea amylose

■,  $r=5$ ; ●,  $r=6$ ; ▲,  $r=7$ ; ▼,  $r=8$ ; ◆,  $r=9$

### 3.2 近红外光谱波段分析

根据茶叶的光谱特征(见图 1)来分析, 全波段  $4\ 000 \sim 9\ 000 \text{ cm}^{-1}$  总共有 1 297 个点, 将这些光谱分为 3 段(I:  $9\ 000 \sim 7\ 500 \text{ cm}^{-1}$ ; II:  $7\ 500 \sim 5\ 500 \text{ cm}^{-1}$ ; III:  $5\ 500 \sim 4\ 000 \text{ cm}^{-1}$ ), 其中 I 波段是近红外谱的二级倍频谱区, 特征

信息极弱,对神经网络模型的建立作用不大;Ⅱ波段是近红外谱的一级倍频谱区,较Ⅰ波段信息稍强,与样品的组成与性质具有一定的相关性,经优化提取特征光谱范围为  $7\,432.3\sim 6\,155.7\text{ cm}^{-1}$ ,共有 332 个点;Ⅲ波段是近红外谱的合频谱区,信息比另两个区要强,从光谱图也可看出,该区中不同茶多酚和茶多糖含量的茶叶其谱线高度有较大差异,包含信息量最丰富,我们选择的特征光谱范围为  $5\,484.6\sim 4\,192.5\text{ cm}^{-1}$ ,共有 336 个点。因此,全波段总共选择了 668 个点来进行神经网络建模。

### 3.3 模型与预测

对于茶多酚和茶多糖的神经网络模型的建立,选择最佳网络结构参数<sup>[4,5]</sup>为 3 层神经网络,其输入层、隐层、输出层神经元数目分别为(8, 4, 1)和(7, 5, 1)。近红外光谱特征范围选择为  $7\,432.3\sim 6\,155.7\text{ cm}^{-1}$  和  $5\,484.6\sim 4\,192.5\text{ cm}^{-1}$ ,共有 668 个点。对已测定的茶多酚和茶多糖含量的茶叶样品,随机抽取 50 个训练样本集进行建模,得到其实验值与由建立模型所预报的预测值的关系图(见图 3),其数据分析的结果为:茶多酚样本的相关系数  $R$  为 0.984 7,误差均方根  $RMSECV$  为 0.460,相对偏差  $RSECV$  为 0.123;茶多糖样本的  $R$  为 0.947 0,  $RMSECV$  为 0.136,  $RSECV$  为 0.224。用所建立的模型对 20 个茶叶预测样本集的茶多酚和茶多糖含量进行预测,得到的结果见表 1,其茶多酚和茶多糖的  $R$ ,  $RMSEP$ ,  $RSEP$  分别为 0.980 4, 0.529, 0.017 和 0.968 2, 0.111, 0.029 8。这些结果表明茶叶中茶多酚和茶多糖各样本的实验值与预测值具有较好的相关性。利用近红

外光谱和建立的人工神经网络模型可预测茶叶茶多酚和茶多糖含量。

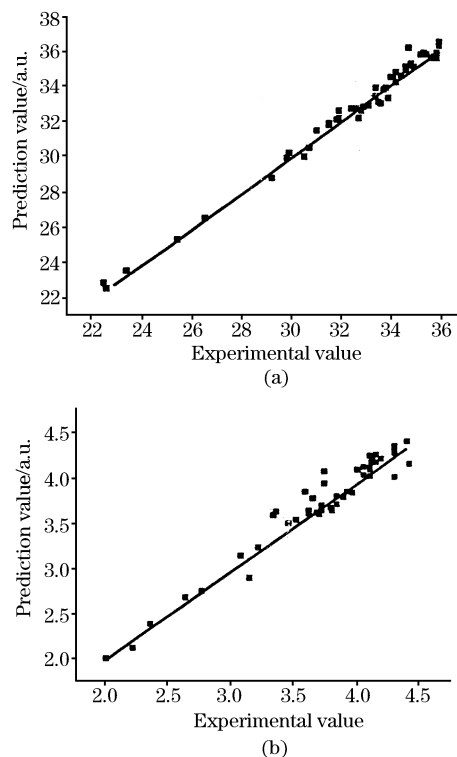


Fig. 3 Relationships between experimental and prediction values of (a) tea polyphenol and (b) tea amylose

Table 1 The experimental and prediction values of tea polyphenol and tea amylose by ANN model

No.	Tea polyphenol				Tea amylose			
	Exp. value	Pred. value	Error	Recovery/%	Exp. value	Pred. value	Error	Recovery/% *
1	24.1	24.5	0.4	101.7	2.35	2.46	0.11	104.7
2	24.6	24.0	-0.6	97.6	2.88	2.79	-0.09	96.9
3	25.3	25.6	0.3	101.2	2.94	2.81	-0.13	95.6
4	26.8	27.3	0.5	101.9	2.98	2.88	-0.10	96.6
5	27.5	26.9	-0.6	97.8	3.06	2.96	-0.10	96.7
6	28.3	29.2	0.9	103.2	3.36	3.29	-0.07	97.9
7	29.3	29.8	0.5	101.7	3.41	3.31	-0.10	97.1
8	29.8	28.7	-1.1	96.3	3.59	3.67	0.08	102.2
9	31.2	31.3	0.1	100.3	3.62	3.53	-0.09	97.5
10	32.1	32.5	0.4	101.2	3.86	3.98	0.12	103.1
11	32.6	32.1	-0.5	98.5	3.97	3.82	-0.15	96.2
12	32.9	33.1	0.2	100.6	3.99	4.17	0.18	104.5
13	33.6	32.9	-0.7	97.9	4.05	3.89	-0.16	96.0
14	33.8	33.6	-0.2	99.4	4.10	4.23	0.13	103.2
15	34.5	35.1	0.6	101.7	4.19	4.27	0.08	101.9
16	34.8	35.0	0.2	100.6	4.21	4.30	0.09	102.1
17	35.2	34.8	-0.4	98.9	4.29	4.34	0.05	101.2
18	35.4	35.8	0.4	101.1	4.31	4.42	0.11	102.6
19	35.6	35.2	-0.4	98.9	4.33	4.42	0.09	102.1
20	35.9	35.4	-0.5	98.6	4.36	4.25	-0.11	97.5

\* ,  $\text{Recovery}(\%) = 100 \times \hat{c}_i / c_i$

## 4 结 论

本实验采用人工神经网络方法,对茶叶的近红外光谱进行了分析,提取了可用于神经网络模型建立的特征光谱,并对神经网络建模参数进行了优化,得到茶多酚和茶多糖近红

外光谱模型的最佳网络结构输入层、隐层、输出层参数。用建立的模型预测了 20 个样本,得到的结果较准确。这些结果在实际应用或进行某成分分析中具有一定参考价值,显示了人工神经网络方法在近红外光谱分霄领域中将有广阔的应用前景。

## 参 考 文 献

- [1] WANG Duo-jia, ZHOU Xiang-yang, JIN Tong-ming, et al(王多加,周向阳,金同铭,等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(4): 447.
- [2] YAN Yan-lu, ZHAO Long-lian, LI Jun-hui, et al(严衍禄,赵龙莲,李军会,等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 20(6): 777.
- [3] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍,袁洪福,徐广通,等). Modern Analysis Technique of NIR(现代近红外光谱分析技术). Beijing: Sinica Petrochemistry Press(北京:中国石化出版社), 2000. 105.
- [4] REN Rui-xue, TANG Zhen, LIU Fu-qiang, et al(任瑞雪,汤真,刘福强,等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2001, 21(4): 521.
- [5] LI Yan, WANG Jun-de, WANG Lian-jun(李燕,王俊德,王连军). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 20(4): 477.
- [6] Zupan J, Gasteiger J. Neural Network and Its Application to Chemistry(神经网络及其在化学中的应用). Hefei: Sinica Science and Technology University Press(合肥:中国科技大学出版社), 2000. 85.
- [7] LU Song-hou, SHI Zhao-peng(陆松候,施兆鹏). The Evaluation and Determination of Tea(茶叶审评与检验). Beijing: Sinica Agricultural Press(北京:中国农业出版社), 2001. 178.

## Studies on ANN Models of Determination of Tea Polyphenol and Amylose in Tea by Near-Infrared Spectroscopy

LUO Yi-fan<sup>1,2</sup>, GUO Zhen-fei<sup>3</sup>, ZHU Zhen-yu<sup>4</sup>, WANG Chuan-pi<sup>1</sup>, JIANG He-yuan<sup>1</sup>, HAN Bao-yu<sup>1\*</sup>

1. Key Laboratory of Tea Chemical Engineering of Ministry of Agriculture, Hangzhou 310008, China
2. College of Chemistry and Environment, South China Normal University, Guangzhou 510631, China
3. College of Biotechnology, South China Agricultural University, Guangzhou 510624, China
4. Basic Medical College of Zhongshan University, Guangzhou 510089, China

**Abstract** The objectives of the present paper were to build the models for the determination of tea polyphenol (TP) and tea amylose (TA) in tea by near-infrared spectroscopy (NIR). According to the range of 7 432.3-6 155.7  $\text{cm}^{-1}$  and 5 484.6-4 192.5  $\text{cm}^{-1}$  of NIR spectra, the models are built for determining the contents of TP and TA in tea with the input layer, hidden layer and node ((8, 4, 1) and (7, 5, 1) respectively) in network structure by the artificial neural network. The correlation coefficient ( $r$ ), the root mean square error of cross validation ( $RMSECV$ ) and the root mean square error of prediction ( $RMSEP$ ) were selected as the indexes for evaluating the performance of calibration models. The results show that  $r$ ,  $RMSECV$  and  $RSECV$  by the model samples for TP and TA are 0.984 7, 0.460 and 0.123, and 0.947 0, 0.136 and 0.224 respectively, and  $r$ ,  $RMSEP$  and  $RSEP$  by the prediction samples for TP and TA are 0.980 4, 0.529 and 0.017, and 0.968 2, 0.111 and 0.029 8 respectively. These indicated that the NIR-ANN models can be used to determine the contents of TP and TA in tea.

**Keywords** Artificial neural network; Near-infrared spectrum; Tea polyphenol; Tea amylose

(Received Nov. 18, 2004; accepted Apr. 3, 2005)

\* Corresponding author