

记技术的发展,各种新的技术不断产生,如果要科学、准确地对植物的遗传多样性和亲缘关系进行研究,还得运用多种分子标记技术,将每种标记的结果和传统的分类结果进行全面的比较和分析,才能获得令人信服的结果。

References:

- [1] Chen X M, Guo S X. Advance in the research of constituents and pharmacology of *Dendrobium* [J]. *Nat Prod Res Dev* (天然产物研究与开发), 2001, 13 (1): 70-75.
- [2] Yang S J. *Plant Biology* (植物生物学) [M]. Beijing: Science Press, 2000.
- [3] Lan M X. *Dendrobium* (石斛) [M]. Beijing: Science and Technology Bibliography Press, 2002.
- [4] Peng R, Li Q S, Li L Y. RAPD based molecular identification of *Dendrobium* species [J]. *J Southwest Agric Univ: Nat Sci* (西南农业大学学报:自然科学版), 2004, 26 (4): 437-440.
- [5] Luo Y M, Ding X Y, Xu L S, et al. Identification of *Dendrobium miniliforme* with PCR technique [J]. *J Huaiyin Teachers Coll: Nat Sci* (淮阴师范学院学报:自然科学版), 2002, 1 (1): 82-86.
- [6] Li T X, Wang J K, Bai Y F, et al. A novel method for screening species-specific DNA probes for species identification [J]. *Nucleic Acids Res*, (核酸研究), 2004, 32 (4): 45.
- [7] Lau D T, Shaw P C, Wang J, et al. Authentication of medicinal *Dendrobium* species by the internal transcribed spacer of ribosomal DNA [J]. *Planta Med*, 2001, 67 (5): 456-460.
- [8] Yu H, He L, Ni N C, et al. Fingerprinting analysis of plant of *Dendrobium* Sw. by AFLP [J]. *Chin Tradit Herb Drugs*. (中草药), 2004, 35 (7): 808-810.
- [9] Tsi Z H. A preliminary study of the orchid genus *Dendrobium* Sw. in China [J]. *Acta Phytotaxonom Sin*. (植物分类学报), 1980, 18 (4): 27-28.

模式识别在白芍质量控制中的应用

冯雪松¹, 刘雅茹¹, 张克荣², 郭晓明¹, 刘俊亭¹

(1. 中国医科大学 药物分析学教研室, 辽宁 沈阳 110001; 2. 沈阳药科大学药学院, 辽宁 沈阳 110016)

摘要:目的 旨在通过通用性强的中药色谱数据特征的抽取和神经网络识别,建立白芍的质量评价模式。方法 首先通过实验获取同一品种不同质量 29 个白芍样本的高效液相色谱数据,然后依照非线性的核主成分分析(KPCA)进行数学特征提取,将取得的压缩数据,输入 BP 神经网络进行学习,运用训练后的网络识别白芍的质量分类。并探讨了模式识别中人工神经网络的数据预处理、网络隐含层数、隐节点数、激励函数和过拟合现象等。结果通过改良后网络训练,已成功地识别白芍药材质量类别(识别率 100%)。结论 非线性特征提取 KPCA 法与人工神经网络结合适用于白芍整体质量分析。

关键词:白芍;质量;高效液相色谱;非线性的核主成分分析

中图分类号:R282.7

文献标识码:A

文章编号:0253-2670(2006)04-0592-04

Application of chemical pattern recognition to quality control of *Radix Paeoniae Alba*

FENG Xue-song¹, LIU Ya-ru¹, ZHANG Ke-rong², GUO Xiao-ming¹, LIU Jun-ting¹

(1. Department of Pharmaceutical Analysis, China Medical University, Shenyang 110001, China;

2. College of Pharmacy, Shenyang Pharmaceutical University, Shenyang 110016, China)

Abstract: Objective With generalization and steadiness, a new evaluation model by Integrating Non Linear Features extraction algorithm with artificial neural networks (ANN) used for pattern recognition of quality control of *Radix Paeoniae Alba* was proposed in this paper. **Methods** The HPLC data from 29 samples with different quality were proceeded with nonlinear kernel principal component analysis (KPCA) and an improved Back propagation algorithm of ANN. The extract characteristics was fed into BP neural networks as input elements for pattern recognition. In the meantime, the processing data, the optimal numbers of hidden layers, the numbers of hidden nodes, excitation functions, and over-fitting, etc. were discussed wholly so that standardization networks was designed without jamming. **Results** As recognition ratio was 100%, the pattern recognition of quality control of *Radix Paeoniae Alba* was established successfully by trained networks and predicted results. **Conclusion** Integrating KPCA algorithm with ANN for pattern recognition of quality control of *Radix Paeoniae Alba* has been proved to be available.

Key words: *Radix Paeoniae Alba*; quality; HPLC; kernel principal component analysis (KPCA)

收稿日期:2005-07-28

作者简介:冯雪松,男,籍贯辽宁锦州,讲师,硕士,毕业于沈阳药科大学药物分析专业,研究方向为中药质量分析。

Tel: 13889161239 E-mail: voncedar@21cn.com, fcedar@sina.com

白芍为毛茛科芍药 *Paeonia lactiflora* Pall. 的干燥根,主产于浙江、安徽、四川等地,味苦、酸,微寒,归肝、脾经。具有平肝止痛、养血调经、敛阴止汗之功能。白芍根主要含有芍药苷、芍药内酯苷、羟基芍药苷、苯甲酰芍药苷和苯甲酰羟基芍药苷等苷类成分。目前,白芍质量标准多单用芍药苷作为白芍的质控指标,难于全面控制白芍内在质量,鉴于中药具有复杂的作用机制和化学成分,属于典型的灰色体系,色谱指纹图谱是全面评价中药质量的有效方法之一,本研究首次将核主成分分析、人工神经网络应用于 RP-HPLC 色谱指纹图谱数据处理,确立了白芍指纹图谱与质量分类之间的数学模型,为全面控制白芍的质量提供了实用方法。

1 算法理论

在中药色谱指纹图谱数据间存在大量线性和非线性关系及耦合交互作用,这对传统模式识别的结论有严重误导,同时“维数灾难”造成算法复杂,计算速度慢。具体对人工神经网络来说,网络结构变得相对臃肿,易陷入局部最小,间接提高了对训练所需最少样本数的要求,而且训练样本相对特征值过少则可能由于“泛化现象”使网络的归纳能力变差,所以必须采取信息提取、压缩,而且数据的特征提取在模式识别中作用具有去除噪音、保证模型的稳健性有重大作用。其中采用非线性数据压缩方法具有更优的非线性类可分性,特征提取算法较多,但每种算法都有其适用范围、适用条件和缺陷。本研究运用非线性核主成分分析和 BP 网络两种算法建立识别模型。

1.1 非线性核主成分分析(kernel principal component analysis, KPCA)^[1]:核主成分分析是通过事先确定的非线性映射函数 φ ,将输入样本数据向量 $A_K(a_{1k}, a_{2k}, a_{3k}, \dots, a_{nk}; m \times n; k=1, 2, 3, \dots, m$ 样本数, n 为特征数),映射到特征空间 $H: \varphi(A_K)$ 中,然后再在该特征空间 H 中对数据进行主成分分析,最大限度地保留原数据所含原始信息,使新变量为原变量的非线性组合,并增强其模式差异,进而运用主因子来研究样本的方法。本研究将其用于原始数据特征提取方法。

1.2 人工神经网络(artificial neural networks, ANN):是一种模拟人脑功能的信息处理系统,以数学网络拓扑结构为理论基础,具有强大的处理信息能力和速度,高度容错能力,并行性信息加工、处理和储存一体化等特点,应用于构效分析、模式识别和过程控制等方面^[2,3]。算法具体步骤略。

2 实验部分

2.1 仪器、试剂与药材样本:日本岛津公司 LC-10AD 液相色谱仪。

芍药苷对照品(批号 0736-200116),白芍对照药材(批号 0905-200106)均购于中国药品生物制品检定所。29 个白芍药材产地见表 1。药材及炮制品通过形态学鉴定为白芍,药材按照条直、均匀,根径大小、质地、粉性程度及有无白心和裂隙为指标;炮制品按规格、色泽是否均匀,杂质数量等指标分为 3 类,Ⅰ类为优质品,Ⅱ类为一般品,Ⅲ类为劣质品(沈阳药科大学孙启时教授鉴定)。

表 1 样本鉴定和模型识别的结果

Table 1 Results of pattern recognition and quality classifications of samples

序号	来源	鉴定类别	预测类别	芍药苷/%
1	中检所	I	I	2.80
2	亳州 I	I	I	2.45
7	徐州	I	I	2.37
5*	杭州	I	I	0.61
6*	河南	I	I	1.42
9	长春	I	I	2.26
11	内蒙	I	I	1.93
14*	湖南	I	I	2.17
15*	安徽 II	I	I	2.26
17	亳州 II	I	I	3.17
19	亳州 III	I	I	1.92
20	徐州炒白芍	I	I	1.83
21	杭州炒白芍	I	I	1.55
24*	杭白芍	I	I	2.55
25*	贵州	I	I	2.20
26	山东曹县	I	I	0.93
27	上海炒白芍	I	I	0.88
28	对照白芍	I	I	2.56
29	四川中江	I	I	3.11
4	兰州	II	II	1.69
8	上海	II	II	1.23
10	山东菏泽	II	II	1.67
13	大连	II	II	1.35
18	兰州炒白芍	II	II	2.29
12*	浙江	II	II	1.25
16	安徽 I	II	II	2.08
3	安徽 III	II	II	1.43
23	亳州炒白芍 1	II	II	1.64
22*	亳州炒白芍 2	II	II	2.33

* 预测集

* predicted sets

2.2 色谱条件:Hypersil C₁₈柱(200 mm×4.6 mm, 5 μm),大连依利特科学仪器有限公司生产;流动相:乙腈-水(12:88, H₃PO₄ 调 pH=3.0);体积流量:0.8 mL/min;检测波长:230 nm;柱温:室温。

2.3 样本制备与分析

2.3.1 对照品溶液的制备:精密称取芍药苷对照品适量,用甲醇配制成 0.25 mg/mL 的溶液。

2.3.2 供试品溶液的制备:精密称取白芍粉末 1.0 g,加入 50%乙醇 10 mL 回流提取 3 次,每次 2 h,滤过,合并各滤液,回收溶剂,残渣用甲醇溶解并定容至 10 mL 量瓶中,摇匀。再精密量取 0.5 mL 置 5 mL 量瓶中,用流动相稀释至刻度,摇匀,微孔滤膜滤过,取续滤液备用。

2.3.3 白芍的指纹图谱:收集测定了不同质量的 29 个白芍样品,共获得 21 个色谱峰,以白芍对照药材为标准确定共有峰,结合不同产地白芍药材,确定其中 12 个共有峰(峰号为 1、2、3、4、5、7、12、13、14、15、20、21)。其中,14 号峰为芍药苷,作为参照物,以其保留时间及峰面积为 1,计算其他各峰的相对保留时间及对应的峰面积比值,将相对峰面积作为图谱特征变量,进行指纹图谱数据分析。样品 11 的色谱图(图 1),对照品芍药苷色谱图(图 2)。

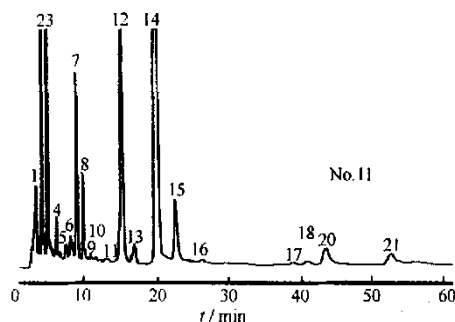


图 1 白芍样品 11 色谱指纹图谱(230 nm)

Fig. 1 HPLC Fingerprints of No. 11 *Radix Paeoniae Alba* sample (230 nm)

白芍样品 HPLC 的各个色谱峰相对面积作为图谱特征变量,对指纹图谱数据进行计算。

3 结果

3.1 对原始数据特征提取同时满足网络输入和最大信息量要求。KPCA 法的前 5 个主因子累计贡献率达 99.57%,因此,KPCA 法将数据由 21 维压缩到 5 维作为输入,隐含节点设 5,形成 $5 \times 5 \times 3$ 的 BP 网络拓扑结构,迭代精度 0.000 01,经多次训练其中 20 个样本,建立成识别模型。

3.2 实验表明,任选 1~n 个白芍色谱指纹图谱数据,经 KPCA 特征提取,输入网络,模型均能按质量等级分类(表 1),正确识别率为 100%。同时测量 29 个白芍样品中芍药苷的量,均符合 2005 年版《中国药典》要求,按 2005 年版《中国药典》要求 I、II、III 类中芍药苷量各有 5/19,3/7,1/3 不符合要求,II、III 类芍药苷量合格率相对较低,同时从另一个角度

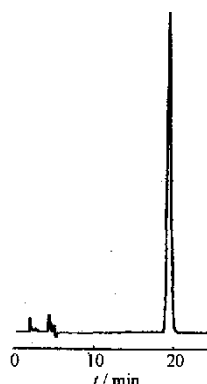


图 2 芍药苷对照品 HPLC 色谱图

Fig. 2 HPLC Chromatogram of paeoniflorin reference substance

说明芍药苷不是代表白芍药材质量的唯一指标。

4 模式识别算法中 ANN 算法的改进与讨论

4.1 训练集、监督集的设计:因为本研究的是药材不同质量优劣识别,所以设计时必须综合考虑训练集中各类样本的数量,选取标准。为能显示出质量间差异最大、质量判别更具说服力,综合标准药材和待测样本,确定指纹图谱上共有峰区间并在此区间选取最强峰,并结合各样本特征指纹峰作为样本化学模式特征。为了

保证样本数据与决策间的高相关性,在满足指纹图谱标准的基础上上尽可能选择同一质量范围内不同质量浓度的样本数据,使训练、监督集具有广泛的代表性,以满足 BP 前馈算法的需要,从而确切描述出数据特征与结论间的因果关系。由于实验样本各质量类别内数量限制,所以选取代表性样本作为训练集,处于训练集样本质量变化范围内的样本作为监督集,以便对网络学习情况进行整体监督。

4.2 数据的预处理:将所有数据处理成符合网络输入、输出要求,即对输入、输出值非线性处理(激励函数本身作为数据压缩、转换函数)已达到标准化的目的,如:反正切函数:

$$y = 2 \times \text{Arctg}(x) / \pi \quad y \in (-1, 1)$$

如此可效果明显,避免每次针对不同问题而对输入值、输出值繁琐变换,使训练好的网络对未知样本预测时更为实用,误差更小,尤其是单一待测样本的原始数据处理、识别。

4.3 网络拓扑结构:本研究采用三层前馈网络。由于网络规模小可能没有较好的兼容能力,网络将不能建立复杂的判决界,过大的网络将失去良好的概括能力,考虑以上优、缺点,采用 KPCA 算法优化隐含节点的方法以确定网络规模。输出采用在模式识别中首选的单输出型,即将属于这一类的样本期望输出设为 0.9,其他设为 0.1,输出节点数对应类别数,以最接近期望输出值所属类别为判别依据,样本所属种类对应的输出节点给出相对于其它输出节点的最大值,为三个输出节点。

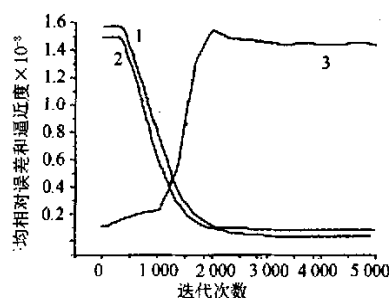
4.4 激励函数的确定:笔者尝试用经典 Sigmoid,

双曲正切等激励函数,并首创用反正切激励函数训练本模式,条件相同时,训练结果以反正切函数最好,收敛最快,精度最高(表2),可能是函数斜率较小,压缩坐标轴,从而摆脱误差函数的平坦区,避免大部分的局部极小值,使算法的收敛速度变快,但有时易发生振荡。

表2 激励函数的比较

Table 2 Comparison of excitation functions

激励函数	平均相对误差	时间(min)/迭代次数
反正切	0.001 835	0.85/5 000
Sigmoid	0.431 175	2.02/5 000
双曲正切	0.022 434	1.31/5 000
平方比(限幅输出)	0.031 689	1.56/5 000
正弦(限幅输出)	0.070 289	1.79/5 000



1-训练集样本平均误差 2-控制集样本平均误差 3-逼近度

1-mean relative error of training set 2-mean relative error of control set 3-approximation degree

图3 ANN逼近度优化

Fig. 3 ANN Optimization by approximation degree

有助于缓解中药色谱指纹图谱客观性和判别主观的矛盾。未来如能制定保证诸如采样的普遍、代表性,制备方法的科学、完善和系统的中药指纹图谱规范要求,结合谱效关系,信息采集、分析自动化等技术,必将极大促进中药现代化的进程。

References:

- [1] Scholkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem [J]. *Neur Comput*, 1998, 10 (5): 1299.
- [2] Zupan J, Gasteiger J. Neural networks: a new method for solving chemical problems or just a passing phase [J]. *Anal Chim Acta*, 1991, 248 (1): 1-30.
- [3] Jin P, Fan J B, Tan Y D. *Principle of Neural Network and Application of Neural Computer* (神经网络与神经计算机) [M]. Chengdu: Southwest Jiaotong University Press, 1991.
- [4] Liu P, Liang Y Z, Zhang L, et al. Artificial neural networks applied in the analysis of chemical data (I) [J]. *Chem J Chin Univ* (高等学校化学学报), 1996, 17 (6): 861-865.

4.5 训练方式:训练网络时先采用累计误差校正算法,即用全部样本都依次输入后计算总的平均误差,校正各连接权值,达到一定精度阈值后,再用标准误差校正算法处理,即每输入单一样本修改一次权值,可以保证网络计算速度与输出精度统一均衡。

4.6 针对神经网络的过拟合现象:由于中药指纹图谱中普遍存在预测样本数与特征数量比低于模式识别的分类要求,此外个别离散数据的存在,网络中不适当的误差精度要求和训练样本设计不合理,均易造成对训练样本的预测能力高于对未知样本的预测能力,则产生过拟合现象。本研究应用刘平等提出逼近度理论^[4]来减少过拟合现象发生,确定最优学习次数及对应的权、阈值参数(图3)。

5 结语

多种识别算法结合,较为有效地避免了单一算法的不足,将其应用于中药指纹图谱的模式识别中,

不同培养基对掌叶大黄毛状根克隆系质量的影响

杨世海¹, 刘晓峰², 果德安³, 郑俊华³

(1. 吉林农业大学中药材学院, 吉林 长春 130118; 2. 吉林亚泰(集团)股份有限公司, 吉林 长春 130031;

3. 北京大学药学院, 北京 100083)

摘要:目的 研究培养基优化和毛状根克隆系对掌叶大黄毛状根生物量和蒽醌产量的影响。方法 选取4个单克隆掌叶大黄毛状根和4种基本培养基,采用统计分析方法,研究影响大黄毛状根生物量积累和5种蒽醌类化合物的产量的因素。结果 不同克隆间大黄毛状根生物量积累存在显著差异,以单克隆毛状根DH7a最高;培养基种类对毛状根生物量的影响存在极显著差异,以WP培养基中毛状根生物量最高;大黄毛状根中的5种游离蒽醌中以大黄酚和大黄素甲醚为主;不同克隆间5种游离蒽醌总产量存在显著差异,以DH7a最高,并且DH7a与B5培养基的组合,5种蒽醌产量最高;克隆和培养基之间互作分析显示克隆DH5a、DH5d与WP培养基,克隆DH7a与B5培养基对于游离蒽醌产量来说是优势组合,而克隆DH5c在4种培养基上5种游离蒽醌产量差异不显著。结论 本研究为利用掌叶大黄毛状根培养系统大规模生产蒽醌类化合物提供依据。