2005 年第 63 卷 化 学 学 报 Vol. 63, 2005
第 24 期, 2216～2220 ACTA CHIMICA SINICA No. 24, 2216～2220

• 研究论文 •

用于中药药品质量快速检测的近红外光谱模糊神经元分类方法

刘雪松 程翼宇*
(浙江大学药物信息学研究所 杭州 310027)

摘要 针对非线性且分类界线模糊的药品质量类别快速测定难题, 将近红外光谱分析与模糊神经网络相结合, 经研究提出近红外光谱模糊神经网络分类方法, 用于计算辨析中药等化学组成复杂药品的近红外光谱模式类别, 从而快速评定这类药品的质量. 以参麦注射液为典型分析对象, 以鉴别其生产厂家这一模式分类问题为例, 考核本文方法, 结果表明, 其分类准确率达到 94.2%, 明显优于经典的 BP 神经网络分类方法(84.6%), 可望用于中药产品质量类别的快速检测与评价.

关键词 药品质量评价; 中药分析; 近红外光谱; 模糊神经网络; 模糊模式分类

Fuzzy Neural Network Classifier for Fast Evaluating the Quality of Chinese Traditional Medicine Products Using Near Infrared Spectroscopy

LIU, Xue-Song CHENG, Yi-Yu*
(*Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310027*)

Abstract To solve the problem of fast identifying the quality sort of chinese traditional medicine products with nonlinear and fuzzy edges of the quality sort, a new method combining near infrared spectroscopy (NIRS) with fuzzy neural network was proposed. The method can differentiate the pattern classification of NIRS of chinese traditional medicine products with complex chemical components, resulting in fast evaluating product quality. An example of distinguishing the manufacturers of Shenmai injection was used to test the performance of the proposed method. The results showed that the classification accuracy reached 94.2%, obviously better than that of classical BP neural network (84.6%). It was verified that the new method could be used for fast evaluating the quality of chinese traditional medicine products.

Keywords quality evaluation of medicine; analysis of chinese traditional medicine; near infrared spectroscopy (NIRS); fuzzy neural network; fuzzy pattern classification

中药等化学组成复杂药品质量的快速检测十分困难, 至今仍缺乏科学合理的仪器分析方法, 形成了中药等天然药物质量分析领域的一个技术盲区. 目前, 中药生产过程中的质量控制及药品市场的有效监控, 迫切需要快速分析的技术手段, 故研究发展先进适用的药品质量快速检测方法具有较高的学术应用价值. 近红外光谱(NIRS)是一种快速、无损和绿色的分析方法, 近年已被引入中药质量分析领域, 陆续被用于中药药效成分的含量测定[1]、中药提取过程分析[2]、天然药物鉴别[3,4]和药材的快速模式识别[5]等, 呈现出有望应用于解决中药产品质量快速检测难题的光明前景. 然而, 这一中药分析新技术还很不成熟, 缺乏全面深入的方法学研究, 尚不能准确地检测、判定中药产品质量类别.

尽管 NIRS 含有丰富的化学信息, 但并不能直接用

* E-mail: chengyy@zju.edu.cn
Received April 21, 2005; revised August 21, 2005; accepted September 16, 2005.
国家自然科学基金重大研究计划重点项目(No. 90209005)及浙江省科技计划重大项目(No. 021103549)资助项目.

于药品的质量鉴别. 质量类别的模式特征信息隐含于纷繁的光谱数据中, 一般需建立模式分类模型, 进而通过模型计算处理数据, 才能得到药品质量分类结果, 这使得建模方法学成为NIRS中药应用技术关键和研究前沿. 许多研究表明, 中药物质体系组成复杂, 其 NIRS 通常存在较强的非线性和各种干扰, 常规的化学模式分类方法难用于此. 神经网络已在非线性复杂化学模式分类系统中得到应用, 并用于建立 NIRS 模式分类模型[6,7]. 然而, 大部分中药产品的质量类别边界较模糊, 经典的神经网络分类模型往往不能准确地区分质量类别, 导致现有的NIRS神经网络分类方法用于评定中药质量类别时, 准确率不太高. 模糊神经网络(fuzzy neural network, FNN)为解决这类难以精确表述类别界限的非线性模式分类问题提供了新的技术途径, 已成功用于化学制剂气体检测的模糊分类[8], 但未见用于产品质量鉴别的NIRS模糊神经网络分类方法的研究报道.

据此, 本文选取参麦注射液为典型分析对象, 以鉴别其生产厂家这一模式分类问题为例, 提出近红外光谱模糊神经网络分类方法. 研究结果表明, 本文方法的分类准确率明显优于 BP 神经网络分类方法, 可望推广用于中药质量类别快速测定.

1　NIR 光谱模糊神经网络分类方法原理

1.1　基本原理及网络结构

先用多重散射校正(multiplicative scatter correction, MSC)和二阶微分法对快速采集到的药品 NIR 光谱数据作预处理, 以消除光谱漂移, 增加光谱数据中低分辨率组分的分辨率并校正基线. 然后, 对经预处理的所有样品NIR光谱数据进行主成分变换, 再将所获的主成分得分数据随机抽取后平均分成 n 组, 并依据多重交叉验证原则组成训练集和预测集, 以便建立分类模型. 各组数据分别经标准化后作为网络输入.

对于预测集样本, 设数据矩阵经 MSC 和二阶导数处理后为 $\boldsymbol{X}_{\text{new}}$, 其主成分得分阵为 $\boldsymbol{T}_{\text{new}}$, 训练集样本载荷阵为 $\boldsymbol{P}_{\text{old}}$, 则 $\boldsymbol{T}_{\text{new}}$ 计算如下[9]

$$\boldsymbol{T}_{\text{new}}=\boldsymbol{X}_{\text{new}}\bullet\boldsymbol{P}_{\text{old}}^{-1} \tag{1}$$

$\boldsymbol{T}_{\text{new}}$ 按下式标准化计算处理后得到 $\boldsymbol{T}_{\text{new}}$ 送入 FNN 计算:

$$\tilde{T}_{\text{new},ij}=\frac{T_{new,ij}-\text{Min}_j}{\text{Max}_j-\text{Min}_j} \tag{2}$$

$T_{\text{new},ij}$ 为 $\boldsymbol{T}_{\text{new}}$ 阵 i 行 j 列数据, Max_j 和 Min_j 分别为训练集主成分得分阵中 j 列最大值和最小值.

模糊神经元分类器的网络结构见图 1. FNN 分成三层: 输入层, 输入节点为 T_i, $i=1, \cdots, p$, 这里 p 为输入节点数, 对应于光谱量测数据的主成分数; 规则和推理隐含层, 规则推理节点 Rule i, $i=1, \cdots, m$, 这里 m 为所建立的推理规则数; 输出层的模糊分类输出 $Y_k\in[0, 1]$, $k=1, \cdots, q$, Y_k 的大小表示归属第 k 类的程度, 当 Y_k 大于某一设定阈值 S_k 时, 可以认为该样本完全属于分类 k, q 为类别数. 输入层到隐含层的节点联接权重为 1, 隐含层规则节点 i 到输出层 k 类节点的联接权重为 a_{ik}, 构成一径向基模糊神经元分类器.

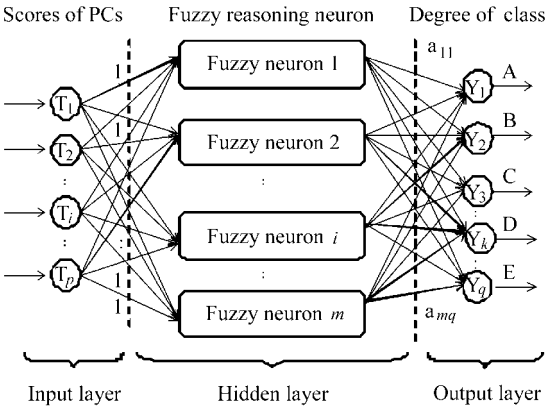

图 1　FNN 网络结构

Figure 1　FNN network architecture

对 FNN 隐含层节点 i 的模糊规则为 R_i: If (T_i is A_{1i}) and ⋯ and (T_p is A_{pi}) then ($Y_1=a_{i1}$) and ⋯ and ($Y_k=a_{ik}$), 其中隶属度函数采用高斯基函数, T_p is A_{pi} 的隶属度计算为:

$$\mu_{A_{pi}}=\exp\left\{-\frac{1}{2}\left[\frac{\left(T_p-C_{pi}\right)^2}{\sigma_{pi}^2}\right]\right\} \tag{3}$$

规则模糊隶属度计算为:

$$\mu_{R_i}=\mu_{A_{1i}}\bullet\mu_{A_{2i}},\cdots,\mu_{A_{pi}}=\exp\left\{-\frac{1}{2}\sum_{n=1}^{p}\frac{\left(T_n-C_{ni}\right)^2}{\sigma_{ni}^2}\right\} \tag{4}$$

FNN 的 k 类归属程度评价输出为:

$$Y_k=\frac{\sum_{i=1}^{m}\left(\mu_{R_i}\bullet a_{ik}\right)}{\sum_{i=1}^{m}\mu_{R_i}} \tag{5}$$

1.2　推理规则生成及网络训练

为生成已知样本的初始经验规则, 将已知训练样本的主成分得分按照各自归属的类别排列. 设属于 k 分类

的总数为 n_k, 由聚类法分别提取主成分得分距离最近的 n_k, n_k-1, n_k-2, …, γ 个主成分得分数据, 经重新排列可生成 $n_k-\gamma+1$ 条规则. 将所生成的规则与 FNN 中分类 k 的连接权重初值设定为训练样本的已知归属度, 而与其它分类的连接权重值设定为 0. 考虑到规则的可信度, 所抽提的最小主成分得分数量宜大于 $n_k/2$, 即 $\gamma > n_k/2$.

规则生成的方法如下: 对第 k 类训练样本的第 m 个主成分得分向量进行聚类分析, 经聚类提取 j 个最近距离主成分得分重新排列后为$[T^k_{n_1m}, T^k_{n_2m}, \cdots, T^k_{n_jm}]^{\mathrm{T}}$, n_1, …, n_j 分别对应各主成分向量在原主成分矩阵中的行数.

$$令 \quad C^k_{mj}=\frac{\sum_{i=1}^{j} T^k_{n_im}}{j} \tag{6}$$

由式(3), 有

$$\sigma^k_{mj}=\frac{\left|T^k_{n_im}-C^k_{mk}\right|}{\sqrt{-2\lg \mu^k_{A_{mj}}}} \tag{7}$$

理论上该得分向量归属 k 分类的隶属度应当为 1, 但由于存在系统误差和其它原因, 实际样本在理论值附近一定范围内按统计规律分布. 设实际隶属度的允许误差范围为:

$$\xi \leqslant \mu^k_{A_{mj}} T^k_{n_im} \leqslant 1, \quad \xi \in (0, 1) \tag{8}$$

故有

$$0 \leqslant \sigma^k_{mj}=\frac{\left|T^k_{n_im}-C^k_{mj}\right|}{\sqrt{-2\lg \mu^k_{A_{mj}}}} \leqslant \frac{\left|T^k_{n_im}-C^k_{mj}\right|}{\sqrt{-2\lg \xi}} \tag{9}$$

取

$$\tilde{\sigma}^k_{mj} \triangleq \mathrm{Max}\left(\frac{\left|T^k_{n_im}-C^k_{mj}\right|}{\sqrt{-2\lg \xi}}\right), \quad n_i=n_1, \cdots, n_j \tag{10}$$

则对不同的 n_i 可以保证相应的 $\mu^k_{A_{mj}} T^k_{n_im} \in [\xi, 1]$.

由上述推导, 对属于 k 类的第 m 个主成分, 其 j 个最近距离主成分的径向基高斯隶属度函数初值可以选为 C^k_{mj} 和 $\tilde{\sigma}^k_{m,j}$.

网络参数的修正采用误差反向传播算法, 以确定的学习速率常数进行网络训练.

2 实验部分

近红外分析仪器采用 ANTARIS 傅立叶变换近红外光谱仪(Thermo Nicolet Corp., USA). 附件配置: 透射检测器, 1 mL 玻璃样品杯, 光程 5 mm. FNN 网络学习、隶属度和规则提取等算法用 Matlab 进行编程(Matlab 6.5, MathWorks Inc.), MSC 和二阶导数处理采用 TQ Analyst 分析软件(Thermo Nicolet Corp., USA).

实验药品参麦注射液来自 5 个生产厂家, 共 120 个样品(其中, 20 个来自生产厂 A, 24 个来自生产厂 B, 24 个来自生产厂 C, 24 个来自生产厂 D, 其余 28 个来自生产厂 E). 样品放入透射样品瓶中进行扫描. 扫描条件: 扫描次数为 64, 分辨率为 8 cm^{-1}, 范围 10000～4000 cm^{-1}. 每个样品扫描 3 次, 取平均谱图. 由于出现水的饱和吸收, 为剔除其对计算的影响, 选择 10000～7300, 7000～5400 和 5200～4000 cm^{-1} 波段数据作为有效数据.

3 结果与讨论

将预处理后的数据中心化, 再进行主成分分析, 其第一主成分对第二主成分作图见图 2. 显然, 仅 A 厂样本能与其他厂样本区分, 其余各厂样本相互间呈交错重叠, 类别界面模糊复杂, 很难精确区分, 因此是一典型的模糊分类问题.

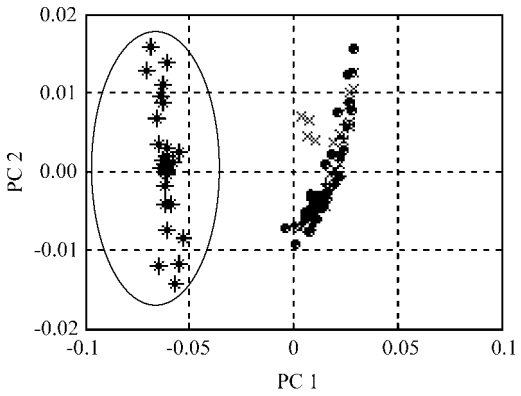

图 2 主成分 1 对主成分 2 关系图

Figure 2 Plot of scores of PC1 versus PC2

*—A; +—B; ×—C; ●—D; —E.

这里, 采用 4 重交叉验证方法考察及比较模型的分类正确率. 对所有样本随机分成 4 组(每组中同一厂家的样本数相同), 取一组数据作为预测集, 其余 3 组数据作为训练集, 共依次轮序进行 4 次独立交叉验证. 图 3 为每次验证时训练集和预测集中各厂样本的分布情况.

对训练集主成分得分阵中每组同厂家的同一主成分得分按最近距离聚类, 选择相近的各分量由前述方法确定隶属度和规则. 本文根据已有学习样本的主成分得分, 对训练样本数据按 5 个不同厂分类, 每类各归纳 4

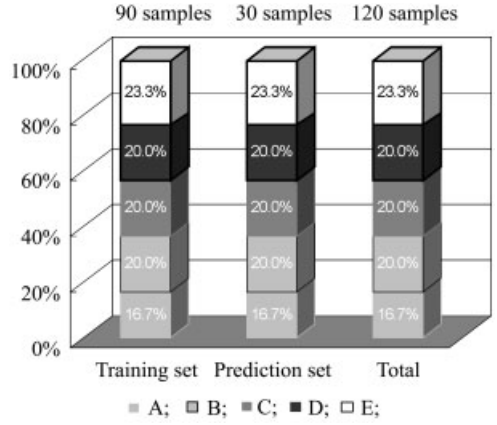

图 3 训练集和预测集样本分布

Figure 3 Data set distribution in every cross-validation

条规则, 共计归纳了 20 条 FNN 的节点规则, 其中分布范围下限值 ξ 取为 0.9.

为选取网络的最佳输入节点数(即主成分数), 这里使用训练集数据用留一交叉验证法计算预测残差平方和:

$$\mathrm{PRESS}=\sum_{i=1}^{n}\left(Y_i-\tilde{Y}_i\right)^2 \tag{11}$$

这里 Y_i 为实际归属度, $\tilde{Y}_i$ 为 FNN 网络计算值. 从 2 个节点开始, 依次增加 1 个节点, 从而得到一系列 PRESS 值, 对应 PRESS 值最小的网络输入节点数即为最佳主成分数. 程序中设定学习速率 λ=0.6, 最小学习误差为 0.01, 每次最大学习步长设定为 1000, 所得最佳主成分数为 10.

图 4 为第 4 和第 5 条模糊推理规则经网络训练后, 神经元节点相应主成分得分高斯基隶属度函数的修正情况. 分析图 4(b)中 Rule 5 的主成分 3 得分隶属度 $\mu_{A_{35}}$, 初始设置中心值为 0.2026, 学习后的中心值为 0.5048, 其学习后中心值修正量为初值的 149%. 实际上, 绝大多数高斯基函数的参数调整量小于 $\mu_{A_{35}}$. 学习前后的数据表明, 由本文方法求出的初始隶属度值与学习所得最优值之间的差距不大. 这说明用聚类均值法求取同类主成分隶属度高斯基函数中心值和用统计分布区间求取宽度值的方法, 可很好地提取出相同类别样本的共性特征, 并能用计算机程序建立可信的模糊规则, 将模糊规则隶属度的初始值界定在一相对较小的学习范围内. 与常规的大范围搜索训练相比, 其学习效率更高, 且波动较小的学习结果也保证了网络参数调整的稳定性.

为与 FNN 分类性能作比较, 另设计一采用误差反向传播算法训练的经典神经网络(BP-ANN), 网络结构与 FNN 相似, 包含输入层、隐含层和输出层, 网络构建

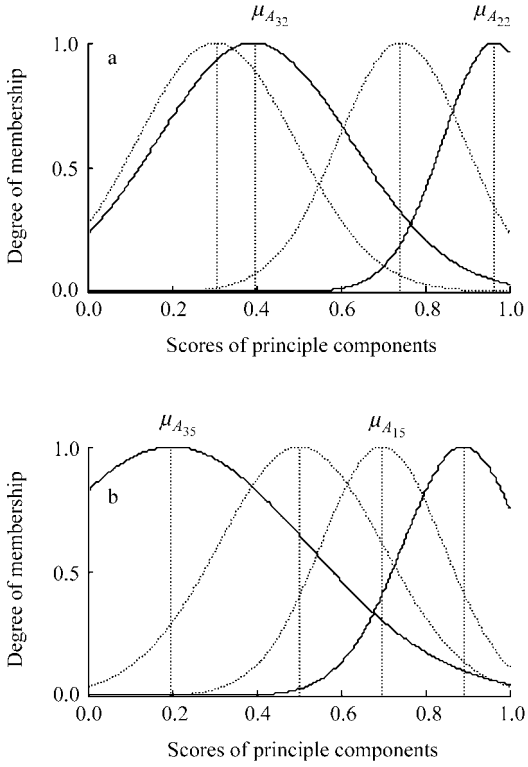

图 4 隶属度函数训练后修正图

(a) 模糊规则 2 中主成分 2 和 3 隶属度函数修正图; (b) 模糊规则 5 中主成分 1 和 3 隶属度函数修正图. ---训练后; —未经训练

Figure 4 Plots of untrained and trained membership function

(a) Gaussian membership function of PC2 and PC3 in fuzzy rule 2; (b) Gaussian membership function of PC1 and PC3 in fuzzy rule 5; --- Trained; — Untrained

和训练采用 Matlab 软件自带的算法和库函数. 其输入与 FNN 的输入数据相同, 隐含层节点函数选择为“tansig”, 输出层传递函数为“purelin”, 训练时设置最大步长为 1000, 训练最小误差为 0.01. 经留一交叉验证法计算 PRESS 得到的最佳隐含层节点数为 18.

当分类阈值 S_k 取为 0.9 时, FNN 与 BP-ANN 对所有样本的分类结果示于表 1. 由于归属程度区间为[0, 1], 计算中对小于 0 和大于 1 的归属度均直接设为 0 和 1. 由表 1 可见, FNN 对未经参与训练的预测集具有较强的外推判断能力, 明显优于经典的 BP-ANN.

表 2 为 FNN 和 BP-ANN 对类别界面严重模糊的 B 厂和 E 厂样本进行两分类的结果比较. 与表 1 的五分类判别结果相比可见, 随着类别数的减少, 两种神经网络的分类正确率均有较大提高, 而 FNN 则显示出更强的经验学习能力和更高的分类正确率. 此外, 在对类别界面较清晰的 A 厂与其他厂样本分类时, 两种神经网络的分类正确率近似, 这就表明 FNN 的优势在于求解模糊分类问题.

表 1 FNN 和 BP-ANN 四重交叉验证分类结果比较

Table 1 Comparison of classification results of FNN and BP-ANN with quadruple cross-validation

	Classifier	Manufacturer					Total	Correction/%
		A	B	C	D	E		
Training set	BP-ANN	60/60	63/72	60/72	63/72	75/84	321/360	89.2
	FNN	60/60	69/72	66/72	66/72	81/84	342/360	95.0
Prediction set	BP-ANN	20/20	16/24	16/24	18/24	22/28	92/120	76.7
	FNN	20/20	20/24	19/24	20/24	25/28	104/120	86.7

表 2 FNN 和 BP-ANN 两厂家四重交叉验证分类结果比较

Table 2 Classification results of FNN and BP-ANN for 2 classes with quadruple cross-validation

	Classifier	Manufacturer		Total	Correction/%
		B	E		
Training set	BP-ANN	66/72	75/84	141/156	90.4
	FNN	72/72	81/84	153/156	98.1
Prediction set	BP-ANN	20/24	24/28	44/52	84.6
	FNN	23/24	26/28	49/52	94.2

4 结论

本文提出的近红外光谱模糊神经元分类方法, 网络结构简单, 节点物理意义明确, 建模训练范围小, 学习效率高, 模糊分类准确率明显优于经典的 BP 神经网络, 可望发展成为一种简便、无损、有效的中药质量类别快速测定方法.

References

1 Liu, X.-S.; Qu, H.-B.; Cheng, Y.-Y. *Chem. Res. Chin. Univ.* **2005**, *21*, 36.

2 Yang, N.-L.; Cheng, Y.-Y.; Qu, H.-B. *Acta Chim. Sinica* **2003**, *61*, 742 (in Chinese).
(杨南林, 程翼宇, 瞿海斌, 化学学报, **2003**, *61*, 742.)

3 Laasonen, M.; Harmia-Pulkkinen, T.; Simard, C. L. *Anal. Chem*. **2002**, *74*, 2493.

4 Tang, Y.-F.; Zhang, Z.-Y.; Fan, G.-Q. *J. Spectrosc. Spect. Anal*. **2004**, *24*, 1348 (in Chinese).
(汤彦丰, 张卓勇, 范国强, 光谱学与光谱分析, **2004**, *24*, 1348.)

5 Woo, Y. A.; Kim, H. J.; Ze, K.-R.; Chung, H. *J. Pharm. Biomed*. **2005**, *36*, 955.

6 Cui, X.-J.; Zhang, Z.-Y.; Ren, W.-L.; Liu, S.-D.; Harrington, P. D. *Talanta* **2004**, 64,

7 Wang, D.; Dowell, F. E.; Ram, M. S.; Schapaugh, W. T. *Int. J. Food Prop*. **2004**, *7*, 75.

8 Gong, J.-W.; Chen, Q.-F.; Fei, W.-F.; Seal, S. *Sens. Actuators, B* **2004**, *102*, 117.

9 Brereton, R. G. *Chemometrics*: *Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons Ltd., West Sussex, **2003**, p. 192.

(A0504211 SHEN, H.; FAN, Y. Y.)