近红外光谱数据分析方法研究

韩 磊,李兴兵,谭跃进
(国防科学技术大学 信息系统与管理学院,湖南 长沙 410073)

摘 要:文中将主成分分析和BP神经网络方法相结合,用于对近红外光谱数据进行预处理和回归分析,较好地解决了近红外分析中的非线性关联问题。实验结果表明,该方法在近红外光谱数据的分析中与传统的化学计量学方法相比有较好的应用效果。

关键词:神经网络;近红外光谱分析;主成分分析法
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2006)05-0001-03

Research on Near Infrared Spectroscopy Data Analysis

HAN Lei, LI Xing-bing, TAN Yue-jin
(College of Information System and Management, National University of Defense Technology, Changsha 410073, China)

Abstract: Principal component analysis (PCA) and back-propagation neural network were used for pretreatment and analysis of the near infrared(NIR) spectroscopy data. The method solved the non-linearity relating problem in NIR analysis. Proved by example, the method used in this article can acquire better result than the traditional chemistry metrology methods in NIR analysis.

Key words: neural network; near infrared spectroscopy analysis; principal component analysis

0 引 言

近红外光谱分析[1]是指利用近红外(NIR)光谱包含的信息对物质特性进行分析的一种技术,主要用于有机物的定性鉴定和定量分析。由于NIR分析技术具有简便、快捷、低成本、无污染以及不破坏样品等优点,因而被广泛应用于石油、化工、医药、烟草等行业。

在近红外分析中,建立稳定的数学模型对于光谱数据分析的准确性至关重要,目前比较成熟的光谱数据分析方法有多元线性回归、偏最小二乘法等。这些建模方法可以很好地反映有机物的光谱数据特征与其特性之间的线性关系,并在实际应用中取得了显著效果。但是,对于物质的光谱特征与其特性之间的非线性关系的建模分析,还需要探讨更好的回归方法。针对这一需求,文中提出采用主成分分析和经典BP神经网络方法对近红外光谱数据进行建模分析,并结合实际数据,验证了该方法的可行性。

1 数据分析方法的基本原理

1.1 近红外分析概述

近红外光谱分析技术最早的应用是对农副产品中的水分、蛋白质、纤维、糖分以及脂肪等成分的分析和测定,并取得了较满意的结果;随着光谱分析仪器和化学计量学方法的不断发展,近红外光谱分析技术也逐渐应用到纺织业、制药、石油工业等领域,并且取得了较好的经济效益。

近红外分析的基本原理是通过测量化合物在近红外谱区的电磁波并对其进行分析而得出化合物的组成成分或者性质等信息。近红外谱区是指介于可见光和中红外谱区之间的电磁波,使人们最早认识的非可见光区域。分子在近红外区的吸收主要由C-H,O-H,N-H和C=O等基团的合频吸收与倍频吸收组成,可以反映出有机物的大量信息。此区的吸收强度低、谱带复杂、重叠严重,无法使用经典的定量或定性分析方法,而必须借助于化学计量学中的多元统计分析、曲线拟合、聚类分析等定标方法将其所包含的信息提取出来,即所谓的"黑箱"分析方法[1]。结合合适的定标模型,可以实现对化合物多种组分或性质的快速分析。

近红外光谱分析方法的应用主要受到硬件和软件两方面因素的制约:有机物样品的近红外光谱的获取,以及样品性质标准值的测定,依赖于精密而准确光谱仪器和各种化学仪器等硬件设施;而对光谱数据的分析,则依赖于分析过程中所使用的数学建模手段和分析算法等化学计量学方法。如前所述,传统的化学计量学方法如多元线性回归、偏最小二乘法等对于建立光谱与化学组成和物理性质之间的线性关联模型具有很好的效果,但是,对于近红外光谱与化合物性质具有非线性关联的情况,传统的化学计量学方法还不能够很好地解决。因此,文中考虑采用人

收稿日期:2005-08-30
作者简介:韩 磊(1980-),男,湖北当阳人,硕士研究生,研究方向为系统管理与综合集成;谭跃进,教授,博士生导师,研究方向为复杂系统理论。

工神经网络方法对近红外光谱数据进行建模分析，以便拟合光谱数据与化合物性质之间的非线性函数关系，为近红外光谱分析技术的进一步发展和应用提供软件方面的支持。

1.2 BP 神经网络算法

人工神经网络[2,3]作为一种由大量简单的处理单元(神经元)广泛互连而形成的复杂网络系统，已经被证明可以以任意精度拟合任意连续函数，具有强大的非线性建模能力。使用人工神经网络进行建模，不需要事先知道模型的具体形式，特别适合于解决复杂的映射问题。在近红外光谱数据分析中，可将样本数据中的光谱信息作为神经网络输入，将化合物的物理或化学性质数据作为输出，对神经网络进行训练，建立起能够映射光谱信息与化合物性质之间关系的神经网络模型，作为近红外光谱数据分析的手段。以文中所举的烟草混合物化学成分的测定为例，可以将代表光谱主要信息特征的主成分作为神经网络输入，将所要测定的总糖、还原糖、植物碱等成分含量作为神经网络输出，来拟合两者之间的线性或者非线性关联。所采用的神经网络结构如图 1 所示。

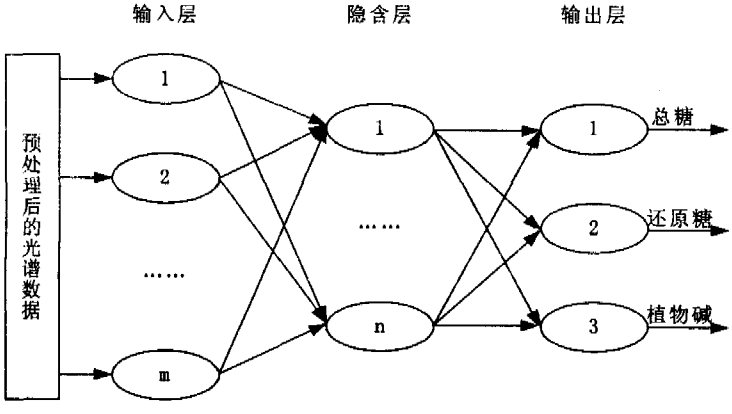

图 1 近红外光谱数据分析中的神经网络结构图

近红外光谱数据分析中运用神经网络方法，需要解决网络的学习能力、推广能力以及抗干扰性等问题。文中提出采用经典 BP 神经网络算法，该算法在网络学习过程中采用反向传播(Back Propagation)学习算法，故称 BP 网络，是一种应用最为广泛的神经网络算法。

BP 学习算法基本原理如下[2]：

设 θ_{pi}^{l} 为 l 层第 i 个神经元的阈值；ω_{ij}^{l} 为 $l-1$ 层第 i 个神经元与 l 层第 j 个神经元之间的权系数；net_{pj}^{l} 表示 l 层第 j 个神经元的输入；o_{pj}^{l} 表示 l 层第 j 个神经元的输出。隐含层或输出层每个神经元的输入为：

$$\mathrm{net}_{pj}^{l}=\sum_{i=1}^{m_{l-1}}\omega_{ij}^{l}o_{pi}^{l-1}-\theta_{pj}^{l} \tag{1}$$

式中：m_{l-1} 为 $l-1$ 层的神经元个数。隐含层或输出层每个神经元的输出为：

$$o_{pj}^{l}=f(\mathrm{net}_{pj}^{l}) \tag{2}$$

式中：f 为 Sigmoid 函数。

$$f(\mathrm{net}_{pj}^{l})=\frac{1}{1+e^{\mathrm{net}_{pj}^{l}}} \tag{3}$$

网络稳定的准则函数为：

$$E_p=\frac{1}{2}\sum_{j=1}^{n}(t_{pj}-o_{pj}^{n}) \tag{4}$$

式中：t_{pj} 为输出层第 j 个神经元的期望输出值。BP 网络权系数 ω_{ij} 修正公式为：

$$\Delta p\omega_{ij}^{n}=\eta\delta_{pj}^{n}o_{pj}^{n-1} \tag{5}$$

$$\Delta p\omega_{ij}^{l}=\eta\delta_{pj}^{l}o_{pj}^{l-1} \tag{6}$$

式中：η 为学习速率；δ 分别为：

当 j 为输出层的神经元，即 $l=n$：

$$\delta_{pj}^{n}=(t_{pj}-o_{pj}^{n})o_{pj}^{n}(1-o_{pj}^{n}) \tag{7}$$

当为隐含层的神经元，即 $l=1,2,\cdots,n$：

$$\delta_{pj}^{l}=o_{pj}^{l}(1-o_{pj}^{l})\sum_{k=1}^{m_{l-1}}\delta_{pk}^{l-1}\omega_{jk}^{l-1} \tag{8}$$

为了加快网络收敛速度，通常权系数 ω_{ij} 修正公式中还需要加一个惯性系数 α，α 为一个常数项，它决定了上一次的权系数对本次权系数的影响：

$$\Delta\omega_{ij}^{l}(k+1)=\eta\delta_{pj}^{l}o_{pj}^{l-1}+\alpha\Delta\omega_{ij}^{l}(k) \tag{9}$$

综上所述，BP 网络的计算一般分为 3 个基本的步骤，即公式(1)和公式(2)中模式的前传；公式(4)中期望输出和实际输出的误差计算，以及公式(5)～(8)中的误差反向传播和权值的修正。

1.3 主成分分析

人工神经网络方法建模的计算量比较大，在应用其进行近红外光谱数据分析时，所建神经网络的输入层数目不能太多，而在近红外分析中所处理的光谱数据往往又是含有大量重叠信息的高维多变量信息，因此，在利用人工神经网络进行建模之前，首先要对近红外光谱数据进行压缩和降维，在保证不丢失光谱主要信息特征的前提下，将高维的光谱数据转化为低维数据，以作为人工神经网络的输入数据。文中采用主成分分析法[4]对光谱数据进行预处理，提取出能够反映光谱主要信息的主成分，作为人工神经网络的输入。

主成分分析法是使用最广泛的线性降维方法之一，该方法概念简单易懂，实现算法高效，因而在许多降维处理中应用都很广泛。比如信号处理中对谱数据的 Karhunnen－Loeve 变换方法，实际上就是主成分分析方法。主成分分析法将方差的大小作为衡量信息量多少的标准，认为方差越大提供的信息越多，反之提供的信息就越少。其基本思想是通过线性变换保留方差大、含信息多的分量，丢掉信息量少的方向，从而降低数据的维数。降维后每个分量是原变量的线性组合，因此，主成分分析方法本质上是一种线性降维方法。其基本原理如下：

考虑数据空间 R^D 中的样本 $\{X_i\}_{i=1}^{N}$，均值 $\bar{X}=\frac{1}{N}\sum_{i=1}^{N}X_i$，协方差阵为 $\boldsymbol{\Sigma}=E(X-\bar{X})(X-\bar{X})^{\mathrm{T}}$，它可以分解为 $\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}$，其中 $\boldsymbol{U}$ 是正交阵，$\boldsymbol{\Lambda}$ 是由协方差阵的特征根构成的对角阵。主成分变换 $Y=\boldsymbol{U}^{\mathrm{T}}(X-\bar{X})$，得到

一个新的数据集合$\{Y_i\}_{i=1}^{N}$,它的均值为 0,协方差阵为对角阵Λ。这样就消除了原来变量之间的相关性。该方法中丢掉那些方差较小的变量,实际上就是将原始数据投影到由前 L 个最大的主成分张成的线性子空间上,从而降低数据的维数。

主成分分析法的计算步骤一般分为以下 4 步[4]:

(1)对原始数据样本集合进行标准化处理。

(2)计算标准化之后的数据矩阵的协方差矩阵,并对其进行正交分解,得出主成分向量。

(3)计算各主成分的累计贡献率,根据要求的贡献率阈值选取主成分。

(4)针对选取的主成分建立主成分方程,计算主成分值。

2　应用实例

2.1　实例描述及原始数据说明

光谱数据中包含了化学成分的大量折射和反射信息,因此可以用来预测某些化学成分的含量。比如,烟草混合物中的总糖、还原糖以及植物碱的含量可以在针对烟草样品摄取的光谱数据中反映出来,因此可以选取合适的波长范围内的烟草混合物光谱数据,通过对其进行分析,来预测烟草中上述 3 种成分的含量。

根据某烟厂利用摄谱仪对某种烟草样品进行的测量结果,选取其中的 5 组实际数据作为样本数据,也就是在分析中所建立的神经网络的输入数据。神经网络的输出变量为总糖、还原糖以及植物碱 3 种成分的具体含量,将通过标准化学方法测定的含量值作为网络的实际输出,对网络进行训练。其具体数值分别如表 1 和表 2 所示。

表 1　波长 4000～9000 范围内的某烟草样品光谱数据

波长(nm) / 样本编号	8994.52259	8990.66560	8986.80860	……	4003.56537	3999.70837
1	0.32875	0.32868	0.32856	……	0.73915	0.73849
2	0.31561	0.31551	0.31542	……	0.73878	0.73806
3	0.31467	0.31463	0.31454	……	0.71518	0.71453
4	0.32180	0.32178	0.32168	……	0.75268	0.75213
5	0.32478	0.32485	0.32481	……	0.69788	0.69729

表 2　表 1 中 5 组烟草样本的总糖、还原糖、植物碱的化学测量值

成分含量 / 样本编号	总糖(%)	还原糖(%)	植物碱(%)
1	33.07	28.41	3.64
2	36.11	29.36	3.12
3	34.92	28.88	2.7
4	34.72	29.15	2.6
5	31.15	27.74	2.03

表 1 中光谱数据的谱图如图 2 所示。

由图 2 可以看出,5 组光谱数据只是在有限的波长范围内才表现出比较明显的差异性,而在大部分波长范围内都存在重复和冗余,因此,有必要对其进行主成分分析,选取其中能够足够反映其差异性的主成分,作为神经网络的输入变量。

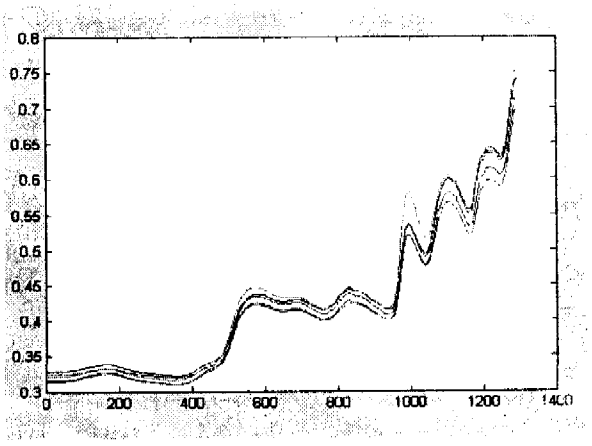

图 2　样本数据光谱谱图

2.2　实验结果

首先,文中通过 MATLAB6.5 编程实现主成分分析法[5],用来对表 1 中的光谱数据矩阵进行预处理。主成分选取的标准是贡献率在 1% 以上的主成分入选,在此条件下对光谱数据进行处理,总共得到 3 个主成分。如表 3 所示。

表 3　对表 1 中数据执行主成分分析的计算结果

主成分 / 样本编号	第一主成分	第二主成分	第三主成分
1	-29.6473	16.8910	8.0328
2	4.4748	-19.5855	5.3526
3	31.4510	-11.3342	1.1892
4	-30.7717	-10.8799	-9.6585
5	24.4932	24.9087	-4.9161

然后采用神经网络仿真工具,拟合光谱数据与 3 种化学成分含量之间的关系。此处采用的人工神经网络共有 3 层。根据上面的主成分分析结果,神经网络输入层共有 3 个神经元,分别对应 3 个主成分;输出层也应该有 3 个神经元,分别对应总糖、还原糖和植物碱的含量;隐含层通过多次仿真试验,根据网络总误差的比较,最后确定为 3 个神经元。网络学习算法采用前面所述的经典 BP 算法,激活函数采用 Sigmoid 函数。

仿真中系统精度为 1×10^{-6},在此精度下网络训练 50000 次后的总误差为 0.002825。

利用训练后的神经网络对 3 种化学成分含量进行预测,得出对应已知的 5 组光谱数据的化学成分含量,预测值与测量值比较如表 4 所示。

表 4　神经网络预测值与测量值的比较

测量值(%)			神经网络预测值		
总糖	还原糖	植物碱	总糖	还原糖	植物碱
33.07	28.41	3.64	0.330323	0.285527	0.363966
36.11	29.36	3.12	0.361617	0.291463	0.312035
34.92	28.88	2.7	0.330197	0.283844	0.236541
34.72	29.15	2.6	0.347569	0.290085	0.260035
31.15	27.74	2.03	0.330182	0.283840	0.236619

由上表可以看出,采用主成分分析方法对光谱数据进

(下转第 87 页)

过融合后的像素 p:$p=\sum_{i=0}^{N}p_iN!/[i!(N-i)!]t^i(1-t)^{N-i}$,$p_i\in R^3$,$0\leqslant t\leqslant 1$。同样,融合系数 t 的取值决定了这 $N+1$ 幅图像的融合效果。

图2所示的六幅图像分别为 $N=2$ 时的三幅原始图像 $P_0(x,y)$,$P_1(x,y)$,$P_2(x,y)$ 以及融合系数 t 分别为0.032,0.5,0.973时的融合图像。

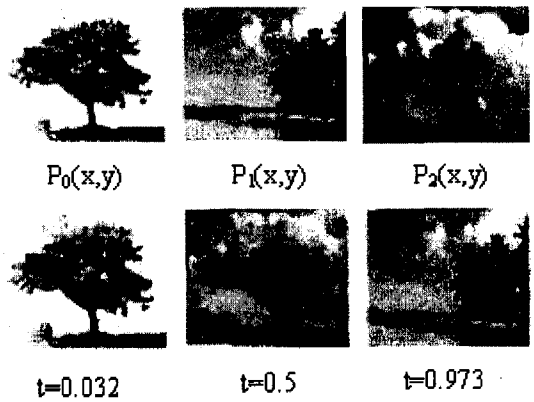

图2　三幅原始图像和取参不同的三幅2阶 Bézier 曲线算法的融合图像

根据融合后图像 p 和 $p_0,p_1,\cdots,p_N$ 中任 N 幅图像就可计算出另一幅隐含的图像,像素恢复算法为:

$$rf_k\text{round}((p-\sum_{i=0,i\neq k}^{N}P_iN!/[i!\quad(N-i)!]t^i(1-t)^{N-i})/(N!/[k!(n-k)!](1-t)^{N-k})) \qquad (7)$$

在实现 N 阶 Bézier 曲线融合算法的时候,可以按照下面的步骤实现 $N+1$ 幅图像(数据文件同理)的融合:

1)在 $N+1$ 幅图像集合中先选出要融合的两幅图像 P_i,P_j,根据需要的隐藏效果选取合适的融合参数 t_1,按照一阶 Bézier 曲线融合算法进行图像融合,融合后的图像为 P。

2)在上述图像集合中删除 P_i 和 P_j,把融合后的图像 P 加入到集合中作为待融合图像。

3)重复上述步骤直到得到最后所需要的宿主图像。

由式(3)和式(4)可知,在计算融合后的像素时 round()函数会引入取整误差。经过分析知道 N 阶 Bézier 曲线融合算法所产生的误差是一阶算法的 N 倍。对于图像,当 N 小于一个临界值时,这些引入的误差并不会影响画质,但是当 N 超过临界值时,将会影响画面质量,同时也会给原始图像的恢复带来困难。

3　结束语

一阶 Bézier 曲线融合算法实现起来较为简单,从可行性、可靠性和高信息隐藏比等几个关键指标来看,该算法做为信息隐藏的一种算法是可选的。但是由于该算法的简单性,所以破解者很容易破解出原始信息,而且该算法的信息隐藏比还可以进一步提高,所以文中提出 N 阶 Bézier 曲线融合加密算法。

信息隐藏是信息安全领域中的热门话题,且研究面相当广泛,像 LSB 算法,还有变换域的算法如 Arnold 变换、Hilbert 曲线变换、Fibonacci 变换等都已得到了广泛应用,而关于 Bézier 曲线的信息隐藏的研究却相对匮乏,由于它的诸多优势,相信它在信息的秘密传输领域会逐渐成为一个研究的热点。

参考文献:

[1] Whitman M E, Mattord H J. 信息安全原理[M]. 北京:清华大学出版社,2004.
[2] Hearn D. 计算机图形学[M]. 北京:电子工业出版社,2002.
[3] Fabien A, Petitcolas P, Anderson R J, et al. Information hiding - a survey[J]. Proc. of IEEE, 1999, 87(7): 1062 - 1078.
[4] 柳葆芳,平西建,邓宇虹. 基于融合的数据隐藏算法[J]. 电子学报,2001,29(11):1445 - 1448.
[5] 郎　锐. 基于一阶 Bézier 曲线的信息隐藏算法[J]. 电脑编程技巧与维护,2004(8):86 - 89.

(上接第3页)

行降维之后,再利用人工神经网络建立校正模型,可以比较准确地拟合出光谱数据与3种化学成分含量之间的关系,其平均拟合精度可以达到0.1%。

3　结　论

文中将主成分分析和 BP 神经网络方法相结合,对一组烟草样本光谱数据进行了拟合。实验结果表明,该方法对于某些烟草成分含量的近红外光谱数据分析能够达到比较高的精度,这在烟草行业具有实际应用价值。与多元线性回归、偏最小二乘等方法相比,神经网络方法的优势在于具有强大的非线性关联能力,可以有效解决光谱数据分析中的非线性校正问题。主成分分析作为一种线性降维方法,可以很好地解决光谱数据从高维空间向低维空间的映射问题,有效解决了神经网络方法在计算量大方面存在的限制。这二者的结合使用,对于近红外光谱分析技术的进一步应用和发展,具有重要的现实意义。

参考文献:

[1] 陆婉珍. 现代近红外光谱分析技术[M]. 北京:中国石化出版社,2000.
[2] 杨荣英. BP 神经网络主成分分析法在交通预测中的应用[J]. 武汉理工大学学报,2002(3):100 - 102.
[3] 王　旭. 人工神经元网络原理与应用[M]. 沈阳:东北大学出版社,2000.
[4] 魏广芬. 基于主成分分析和 BP 神经网络的气体识别方法研究[J]. 传感技术学报,2001(4):41 - 47.
[5] 成卫青. 应用主成分回归分析评估企业经济效益[J]. 微机发展,2003,13(7):29 - 31.