第 12 卷 第 5 期 物 理 化 学 学 报 Vol.12, No.5
1996 年 5 月 ACTA PHYSICO-CHIMICA SINICA May, 1996

Cm I 奇宇称光谱能级的模式识别研究

曹晓卫 刘洪霖 陈念贻
(中国科学院上海冶金研究所，上海 200050)

摘要 应用新的模式识别方法 PCA-BPN(Principal Component Analysis-Back Propagation Network) 指认 Cm I 奇宇称未知能级，支持了前人应用传统的 KNN(K Nearest Neighbors) 等模式识别方法及对传神经网络方法 (Counter Propagation Network,CPN) 对大部分谱线的指认，进一步确认了这些组态的归属；鉴别了 KNN 等与 CPN 不同的预报结果，纠正 CPN 的某些错误分类，并以可视非线性映照分类器加以佐证.

关键词： Cm I 奇宇称光谱， 能级分类， 模式识别， PCA-BP 神经网络， 非线性映照

原子光谱的电子组态通常是根据谱线的能级、强度、同位素位移、塞曼效应等测量数据进行确定，或者应用量子理论计算来指认. 但是，由于光谱的复杂性，上述的实验观测和理论计算难于确定某些高激发态的原子光谱所属的电子组态. 另一方面，模式识别方法在化学上已经得到广泛的应用，在化学体系分类、材料制备的计算机辅助设计以及化学过程的优化等方面，都取得了相当的进展 [1-3].

早在七十年代末，Peterson 等就用当时流行的化学模式识别的 KNN 等方法，借助于由实验测定的 Cm I 、 H I 、 U I 和 U II 等原子光谱数据，指认它们中尚未确定的高激发态光谱线的电子组态，取得了一些积极的效果 [4-7]，表明化学模式识别方法适用于该类问题的研究. 1978 年，Peterson 等人依据 Worden 和 Conway 的光谱实验数据 [9]，用 KNN 等方法指认了 Cm 部分未知能级的电子组态 [5]. 但有相当一部分能级的组态难以确认. 近几年来，人工神经网络算法也已成为有效的模式识别方法. Peterson 应用神经网络算法中的对传网络方法 (CPN)，重新研究了上述 Cm I 奇宇称未知原子光谱，得到了许多新的结果 [8]. 正如表 1 所示，CPN 计算大多肯定了 KNN 等方法的结果，对于 KNN 等无法确定归属的谱线也已给出明确指认.

然而，CPN 计算也有与 KNN 等方法相反的结果. 例如表 1 的第四个能级为 24900.55cm^{-1} 的谱线，KNN 方法认为是 $5f^76d^27s$ 组态，而 CPN 认为是 $5f^87s7p$ 组态. 由于无论是 KNN 等模式识别方法还是 CPN 方法，都属于经验的计算方法，目前尚不能从物理理论的角度对此类相矛盾的结果加以判别孰是孰非. 另外，CPN 方法对 KNN 等方法无法明确判定的谱线所作的指认，Peterson 也担心不甚可靠 [8]. 其次，CPN 算法是较早的神经网络算法，弱点较多，其中最为严重的是，当学习样本数较少时计算可靠性差 [10]. 不巧，这里所涉及的 Cm I 原子光谱奇宇称电子组态的课题恰恰存在样本数少的弊病，例如 $5f^76d^3$ 组态的学习样本才有 4 个，导致在该情形下 CPN 方法的预报结果难以置信. 同时，测试 CPN 算法对于 Cm I 奇宇称已知能级学习预报的准确率仅为 90. 9%[5]，何况这一测试还不是很严格. 实际上，Cm I 的原子光谱仍需进一步的研究.

1995-09-18 收到初稿， 1995-11-27 收到修改稿. 联系人：刘洪霖. * 国家自然科学基金资助项目

十多年来，化学模式识别方法发展了许多新算法，近来广泛应用的 PCA(Principal Component Analysis) 和 NLM(NonLinear Mapping) 方法表明它们比 KNN 等早期的算法可靠. PDP 小组提出新的有监督的神经网络算法 — 反向传播网络 (BPN, Back Propagation Network)[11]，已被广泛应用于模式识别研究. BPN 方法应用于模式识别要较 CPN 方法更为可靠，而且，BPN 已有了新的发展 —PCA-BPN[12], 它消除了网络的输入噪声，提高了网络的训练速度和预报准确率. 并且，只要总的已知样本数足够大，PCA-BPN 方法的可靠性就不依赖于其中某一特定类的已知样本数. 此外，借助于 PCA-NLM (Principal Component Analysis-NonLinear Mapping) 方法可形成一可视分类器以直观地显示所研究的未知样本其周围已知样本的分布这一功能，对 PCA-BPN 的预报结果加以佐证. 因此，应用 PCA-BPN 代替 CPN 重新研究 CmI 的高激发态光谱的电子组态并结合模式识别的 PCA-NLM 算法，指认某些尚有争议的能级归属，应是有意义的工作.

本文将介绍应用 PCA-BPN 和 PCA-NLM 方法，研究 CmI 的高激发态光谱电子组态得到的新结果. 这是 PCA-BPN 和 PCA-NLM 方法首次应用于原子光谱未知组态能级的指认研究. 计算过程中所用到的计算机程序是我们根据有关的原理和算法自行编制的.

1 方法

1.1 PCA-BPN

一般地，BPN 通过构筑具有一定拓扑结构的前向网络、以一组网络连接权值的形式来描述各变量与目标间的非线性关系. 从广义的角度讲，它可视为一类特殊的非线性回归分析，并且要尽量减少独立变量的数目以避免“过拟合”现象、求得准确和可靠. 因此，去除冗余变量十分必要. PCA 通过对原始变量进行主成分变换获得相互正交的独立变量，并且选取对应于本征值较大的若干主成分作为网络的输入，从而达到删除冗余变量的目的.

在此，简要地阐述 PCA 有关的内容，详细内容可参阅文献 [13].

根据 PCA 的原理 [13] 可以得到

$$T = XP \tag{1}$$

式中 X 是已标准化的样本矩阵 (由带有 m 个特征的 n 个样本构成), T 为 X 在主成分空间的投影即得分矩阵. P 则为样本集协方差矩阵的本征矢矩阵 (也称主成分矩阵). 求得的 T_k 所包含的统计噪声小于 $T_{k+1}(k=1, 2, \cdots, m\text{-}1)$. 由 PCA 原理可知，其主成分包含的信息可靠性从第一主成分到最后一个依次降低. 噪音大的主成分被剔除后，可提高信息处理结果的可靠性. 在 PCA-BPN 中，只有包含很小噪声的一部分 T 作为 BPN 的输入，而非原始变量 X 或全部的 T. 这是 PCA-BPN 与 BPN 的主要区别. 本质上，是以一些原始变量的正交函数作为神经网络的输入元，滤除了部分噪声、改善了输入信息. 而且，由于 BP 网络输入层节点个数的减少，统计可靠性得以提高.

1.2 PCA-NLM

NLM 方法最初是由 Sammon 提出的，后由 Kowalski 等加以改进的一种模式识别方法 [13]. NLM 方法的原理可由下式描述：

$$E = (1/\sum_{i>j}^{N} d_{ij}^{*}) \sum_{i<j}^{N} [(d_{ij}^{*} - d_{ij})^2 / d_{ij}^{*}] \tag{2}$$

式中 d_{ij}^* 为原始高维空间中样本 x_i 与 x_j 间的距离，d_{ij} 为映射后在二维平面上两相应投影点之间的距离．N 代表样本数．通过使误差函数 E 达到极小或尽可能地小，获得 X 在该条件下在二维平面上的非线性映照．通常高维空间的信息是直接应用原始的特征参数，考虑到 PCA 所抽取的前 n 个主成分包含的可靠信息更集中，因此，类似于前述的 PCA-BPN，在 PCA-NLM 中，以部分带有较小噪声的主成分 T 作非线性映照，投影到二维平面．平面坐标 Y_1 和 Y_2 是在使误差函数 E 达到极小时主成分 T_k 的非线性组合．在 PCA-NLM 中，(3) 式中 d_{ij}^* 为 $s(s<m)$ 维 PCA 空间中样本 x_i 与 x_j 间的距离，d_{ij} 则为相应的二维平面上两样本投影点间的距离．因此，在 PCA-NLM 方法中，应用 PCA 方法所抽取的前 n 个主成分构成子空间，先对原始的高维空间降维，使被作非线性投影的信息正交，然后再作 NLM，投影到二维空间，构成可视分类器．

作出二维的 NLM 的分类图后，根据下式定义的非线性空间势能函数式中 n_1 、n_2 分别代表两类已知样本的数目，r_{iq}(或 r_{jq}) 表示样本 i(或 j) 到某一点 q 的距离，w 是一常数 [13]，计算图上各点势能，获得两类学习样本势能分界线．依据未知样本在图上的分布位置确定其所属的类型．

$$Z(q)=\frac{1}{n_1}\sum_{i}^{n_1}\frac{1}{1+wr_{iq}}-\frac{1}{n_2}\sum_{i}^{n_2}\frac{1}{1+wr_{jq}} \tag{3}$$

进一步定义图上位置 q 的类型值 $C(q)$：若 $Z(q)>0$，则 $C(q)=1$；反之，若 $Z(q)<0$，则 $C(q)=2$．这些定义为样本空间分类提供了近似判据．因此，可依据未知样本 q 的 $Z(q)$ 值判定其类别．

在此，学习样本集 X 由 CmI 的 55 个四个特征 (光谱能级 (cm^{-1}) 、总角动量量子数 J 、g 因子和同位素位移 S) 均已知的能级组成．这 55 个已知能级分属于四种不同的电子组态，其中包括 30 个 $5f^76d^27s$ 组态，10 个 $5f^76d7s^2$ 组态，4 个 $5f^76d^3$ 组态，11 个 $5f^87s7p$ 组态．为便于问题的研究，把 55 个已知样本分成如下三个体系进行 PCA-BP 神经网络的分类学习和预报 (均以前一组态样本为一类，后一组态样本为二类)：(1) $5f^76d^27s$ 与 $5f^87s7p$；(2) $5f^76d^27s$ 与 $5f^76d7s^2$；(3) $5f^76d^27s$ 与 $5f^76d^3$．

将 X 作主成分变换后可得到四个主成分矢量和样本集在其上的投影 $T_k(k=1\sim4)$．其中只有 T_1，T_2 和 T_3 用于 PCA-BPN 及 PCA-NLM 的计算．

此时，式 (2) 中的 d_{ij}^* 为：

$$d_{ij}^*=\sqrt{\sum(T_{ik}-T_{jk})^2} \tag{4}$$

上式中是对变量 T_{ik} 从 $k=1$ 到 3 求和 (即在此以 PCA 的前三个主成分 T_{ik}，$k=1, 2, 3$，构成子空间)，而不是用 X 直接计算．

在此所采用的 BP 网络的结构为：输入层具有四个输入节点 (T_1, T_2, T_3 和一偏置)，隐层为具有三个隐节点的单隐层，输出层为单输出．以样本的类型 (即电子组态) 为输出目标 (两类样本的类型值定义为 1 和 2)，预报值小于 1.3 为一类，大于 1.7 的为二类，介于其间的则为不可分．

2 结果与讨论

我们将 PCA-BPN 和 PCA-NLM 方法应用于 CmI 奇宇称未知能级的分类．CmI 奇宇称组态未知的十二条谱线的特征变量 [9] 见表 1 所示．

通过已知的 55 个样本按上述三个体系分类学习，获得相应的三个 PCA-BP 网络. 用所得到 PCA-BP 网络对未知能级的组态进行预报，实际预报结果是所有的未知样本的预报值均未落在 [1.3, 1.7] 这一区间上. 此外，值得注意的是，对已知样本以留一法 (leave-one-out) 用所获得的 PCA-BP 网络学习预报的准确率为 100%.

PCA-BP 网络对未知能级的预报结果见表 1 所示. 同时，由 PCA-NLM 方法可获得相应的分类映射图，对于有争议的两个未知能级的 PCA-NLM 结果示于图 1 、 2. PCA-NLM 基于 $Z(q)$ 值判定的未知能级的类别列于表 1.

从表 1 可以看出，对于所有未知样本，PCA-BPN 和 PCA-NLM 给出了一致的预报结果. 与 Peterson 等报导的结果相比较，获得如下一些令人感兴趣的结果.

表 1 几种模式识别方法对 CmI 奇宇称未知谱线的指认

Table 1 Configuration predictions of several pattern recognition techniques for unknown odd-parity energy levels of CmI

No	1	2	3	4	5	6
E/cm^{-1}	22640.04	23282.58	23299.58	24900.55	25518.80	25878.11
J	5	4	5	3	6	5
g	0.98	1.176	0.968	1.54	1.51	1.49
S	0.00	0.00	0.00	−0.40	−0.39	−0.35
KNN etc.	$5f^76d^27s$ 或 $5f^76d7s^2$	$5f^76d^27s$ 或 $5f^76d7s^2$	$5f^76d^27s$ 或 $5f^76d7s^2$	$5f^76d^27s$	$5f^76d^27s$	$5f^76d^27s$
CPN	$5f^76d7s^2$	$5f^76d7s^2$	$5f^76d7s^2$	$5f^87s7p$	$5f^76d^27s$	$5f^76d^27s$
PCA-BPN	$5f^76d7s^2$	$5f^76d7s^2$	$5f^76d7s^2$	$5f^87s7p$	$5f^76d^27s$	$5f^76d^27s$
PCA-NLM	$5f^76d7s^2$	$5f^76d7s^2$	$5f^76d7s^2$	$5f^87s7p$	$5f^76d^27s$	$5f^76d^27s$
No	7	8	9	10	11	12
E/cm^{-1}	28487.41	28634.99	28880.03	28989.06	31104.82	31167.95
J	4	4	5	4	4	4
g	1.328	1.731	1.46	1.648	1.485	1.759
S	−0.18	−0.08	−0.58	−0.25	−0.41	−0.12
KNN etc.	$5f^76d^27s$	$5f^76d^27s$ 或 $5f^76d7s^2$	$5f^76d^3$	$5f^76d^27s$	$5f^76d^3$ 或 $5f^76d^27s$	$5f^76d^27s$
CPN	$5f^76d^27s$	$5f^76d7s^2$	$5f^76d^3$	$5f^76d^27s$	$5f^76d^27s$	$5f^76d^27s$
PCA-BPN	$5f^76d^27s$	$5f^76d^27s$	$5f^76d^3$	$5f^76d^27s$	$5f^76d^3$	$5f^76d^27s$
PCA-NLM	$5f^76d^27s$	$5f^76d^27s$	$5f^76d^3$	$5f^76d^27s$	$5f^76d^3$	$5f^76d^27s$

对于 5 、6 、7 、9 、10 、12 这六个未知样本，KNN 等方法、CPN 方法和 PCA-BPN 、PCA-NLM 方法给出相一致的预报结果，增加了这些组态归属的可信度. 对于 1 、 2 、 3 这三个未知样本，KNN 等模式识别方法对其电子组态无法作出准确的判定，CPN 方法判定为其均属于 $5f^76d7s^2$ 这一组态 (KNN 认为的两种可能组态之一), PCA-BPN 及 PCA-NLM 方法支持了 CPN 方法的预报. 就 PCA-BPN 及 PCA-NLM 与 CPN 的预报结果相比较而言，除第 8 与第 11 两个未知样本外，对其余十个未知样本组态归属的预报结果一致.

对于第四号未知样本，KNN 等方法指认为 $5f^76d^27s$ 组态，而 CPN 则指认为 $5f^87s7p$ 组态. PCA-BPN 预报其属于 $5f^87s7p$ 组态. 从 PCA-NLM 结果同样可以看出该未知样本位于势能分界线 $5f^87s7p$ 样本区域一侧，属于该组态. 由此，我们的 PCA-NLM 和 PCA-BPN 的

指认都和 CPN 一致，否定了 KNN 等方法的判别.

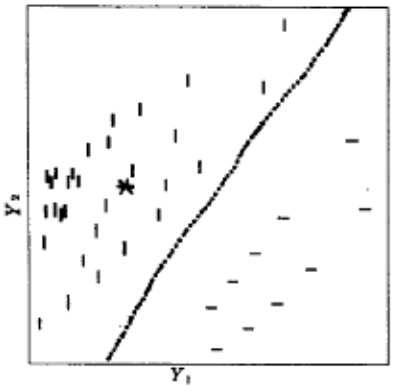

图 1 $5f^76d7s^2$ 与 $5f^76d^27s$ 的 PCA-NLM 图

Fig.1 The PCA-NLM for $5f^76d7s^2$ remarked as −, and $5f^76d^27s$ as |. The * is the 8th unclassified (unknown) configuration to be predicted

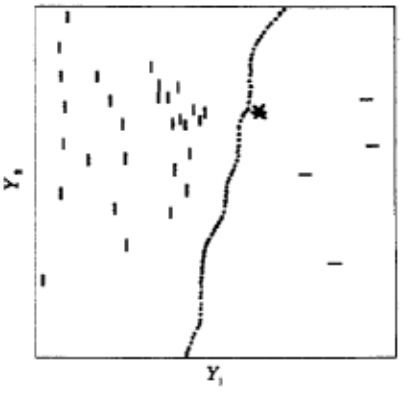

图 2 $5f^76d^3$ 与 $5f^76d^27s$ 的 PCA-NLM 图

Fig.2 The PCA-NLM for $5f^76d^3$ remarked as −, and $5f^76d^27s$ as |. The * is the 11th unclassified (unknown) configuration to be predicted

值得注意的是，对于第八号未知样本，KNN 等认为其属于 $5f^76d7s^2$ 或 $5f^76d^27s$，无法准确指认其电子组态的归属. CPN 指认为前者，但 PCA- BPN 给出的预报结果却是后者. PCA-NLM 结果 (图 1) 显示该未知样本 (*) 位于 $5f^76d^27s$ 类样本聚集区域中心，远离势能分界线，因此应认定为 $5f^76d^27s$ 组态，从而佐证了 PCA-BPN 的指认是正确的.

同样地，对于第十一号未知样本，KNN 等方法认为其属于 $5f^76d^27s$ 或 $5f^76d^3$，也无法准确指认其所属的组态. CPN 预报为前者，而 PCA-BPN 却预报为后者. 在相应的 PCA-NLM 的分类图 (图 2) 上，该未知样本位于分界线 $5f^76d^3$ 样本区域一侧，属于 $5f^76d^3$ 组态，两种方法的指认是一致的，否定了 CPN 的指认. 从 PCA-NLM 图可知，该未知样本处于势能分界线附近，而且 $5f^76d^3$ 组态代表样本少，只有 4 个. 也许，这是造成 CPN 方法误报、KNN 等无法对其所属组态明确指认的原因之一.

综上所述，PCA-BPN 和 PCA-NLM 方法是对原子光谱未知谱线进行分类指认的更为完善和可靠的方法，在原子光谱研究中有着广阔的应用前景.

参 考 文 献

1 Chen N Y.*J. Analytica Chimica Acta*, **1988, 210:**175

2 Liu H L, Chen N Y, Lu W C, Zhu J W. *Analytical Letters*, **1994, 27:** 2195

3 Liu H L, Chen Y, Chen N Y. *J. Chemometrics*, **1994, 8:** 439

4 Peterson K L, Parsons M L. *Phys. Rev.*, **1978, A17:** 261

5 Peterson K L, Anderson D L, Parsons M L. *Phys. Rev.*, **1978, A17:** 270

6 Lewis R V, Peterson K L. *Phys. Rev.*, **1987, A35:**1119

7 Lewis R V, Peterson K L. *Phys. Rev.*,**1988, A38:** 3773

8 Peterson K L. *Phys. Rev.*, **1990, A41:**2457

9 Worden E F, Conway J G. *J. Opt. Soc. Am.*, **1976, 66**:109
10 Hecht-Nielsen R. *Applied Optics*, **1987, 26**:4979
11 Rumelhart D E, Hinton G E, Williams R J.*Nature*, **1986, 332**:533
12 Gemperline P J, Long J R, Gregoriou V G.*Anal. Chem.*, **1991, 63**:2313
13 Varmuza K.Lecture Notes in Chemistry, Vol. 21: Pattern Recognition in Chemistry, Springer-Verlag Press, 1980

Study on the Curium I Odd-Parity Energy Levels Using Pattern Recognition Techniques

Cao Xiaowei Liu Honglin Chen Nianyi
(Shanghai Institute of Metallurgy, Chinese Academy of Sciences, Shanghai 200050)

Abstract A new pattern recognition technique PCA-BPN(principal component analysis-back propagation network) has been used to assign the unknown electronic configurations of odd-parity energy levels of the first spectrum of curium (Cm I). The obtained results show that (1) most previous predictions given by KNN(K nearest neighbours) and CPN(counter propagation network) are further confirmed;(2) several energy levels, which could not be clearly assigned by KNN etc., are predicted to be in good agreement with the assignments of the CPN;(3) two energy levels which were wrongly predicted by the CPN are now corrected using the PCA-BPN and the new assignments are supported by the traditional pattern recognition technique, PCA-NLM(principal component analysis-nonlinear mapping).

Keywords: CmI odd parity spectrum, Classification of energy levels, Pattern recognition, PCA-BP neural network, Nonlinear mapping