第21卷 第3期　　高能物理与核物理　　Vol.21, No.3
1997年 3月　HIGH ENERGY PHYSICS AND NUCLEAR PHYSICS　Mar., 1997

用神经网络分辨原初宇宙线成分*

梁化楼　解　卫　任敬儒　王泰杰　戴贵亮

(中国科学院高能物理研究所　北京　100039)

1996-02-05收稿

摘　　要

尝试在高山乳胶室实验中用神经网络的方法区分超高能区原初宇宙线当中的质子和原子核，对模拟数据的分析结果表明，当族事例观测能量大于500TeV时，对质子和原子核的分辨率均能稳定在80%附近；而当族事例观测能量在100TeV和500TeV之间时，对质子和原子核的分辨率均大于70%.

关键词　神经网络，遗传算法，原初宇宙线成分，高山乳胶室.

1　引　　言

原初宇宙线在小于10^{14}eV能区的成分可以通过高空实验直接测定，而在大于10^{14}eV能区，始终没有直接测量的确定结果. 高山乳胶室实验主要是通过对超高能族事例的研究，来探寻超高能核作用的规律. 由于初级成分不清楚，又经过很长距离的空气级联传播，导致分析中有许多不确定因素，对作用机制难以得出明确的结论. 因此，确定超高能区初级成分有相当重要的意义.

近年来，人工神经网络(ANN)以它信息存储的分布性、信息处理的并行性、容错性、自组织和自学习等优良特性，引起了众多领域研究者的兴趣. 在高能物理实验中，神经网络的模式识别方法也日益增多，并已被成功地用于加速器实验中带电粒子的径迹重建[1]和夸克与胶子产生的喷注的区分[2]. 本工作尝试在乳胶室实验中用神经网络分析族事例来鉴别原初质子和原子核，并取得了比较好的结果.

我们采用一种新型的网络——遗传级联相关网络(GCCA)对HDSQI模型[3]的模拟数据进行分析. 结果表明，当族事例观测能量大于500TeV时，对质子和原子核的分辨率均能稳定在80%附近；而当族事例观测能量在100TeV和500TeV之间时，对质子和原子核的分辨率均大于70%.

2　神经网络模型

神经网络模型一般要考虑三个因素，即人工神经网络的拓扑结构、基本单元(神经

* 国家自然科学基金资助.

元)的特征和网络训练的学习算法. 本工作采用的是将遗传算法[4]和级联相关算法[5]结合而成的遗传级联相关算法[6]. 其主要特点是能够根据问题的需要自动设计和优化网络结构. 目前已被成功地应用于北京谱仪(BES)上 J/ψ 衰变中末态强子的鉴别和各种复杂问题的求解[7]. 该算法概述如下:

(1) 网络初始化. 生成仅由输入和输出组成的最小网络.

(2) 训练输出层. 对输出层权重编码, 随机生成初始群体, 用遗传算法(GA)进行训练. 若学习完成则停止, 否则, 经过一定的"忍耐"进化次数, 网络训练的误差没有明显变化或达到了限定的最大进化次数时, 转向用 Windrow-Hoff[8] 或 δ 规则等任何单层网络训练算法来训练网络(本工作用 Quickprop 算法[9]), 若学习完成, 则停止, 否则, 经过一定"忍耐"次数训练, 训练误差没有明显变化或达到了限定的最大训练次数时, 则转到下一步.

(3) 初始化候选点. 随机生成初始化群, 群体中每一个体表示该候选节点与网络中所有输入节点及所有已生成隐节点间连接权重的编码.

(4) 训练候选节点. 用 GA 优化初始化后的候选节点, 再用 Quickprop 进行训练, 经过一定的"忍耐"次数, 如果候选节点的输出和网络误差间的相关值没有明显提高或达到了规定的最大训练次数时, 转向下一步.

(5) 安装新的隐节点. 选出适合度最大的节点, 连入网络.

(6) 冻结新生成的隐节点的输入权重, 初始化它与输出节点间的连接权重, 返回第二步.

GCCA 的基本思想是利用 GA 的全局搜索长处结合级联相关算法的生长特点, 来防止网络在训练过程中陷入局部极点值, 同时, 在网络生长过程中由 GA 对网络的初始权重进行"粗选"以减少训练结果对初始权重的敏感性, 而后由神经网络方法进一步细调以精确定位解的位置, 从而达到取长补短的效果.

3 GCCA 对原初宇宙线成分的鉴别

在乳胶室实验中, 原初成分的信息主要被包括在族事例当中, 因此, 采用 HDSQI 模型的铅乳胶室模拟族事例, 对 GCCA 进行了训练和测试, 为消除电磁级联过程的影响, 预先对事例进行了退级联[10]处理.

3.1 HDSQI 模型简介

HDSQI 模型是富士和甘巴拉实验组为分析乳胶室实验数据而建立的. 该模型模拟了超高能宇宙线粒子穿过大气到达乳胶室的全部过程, 并且非常好地再现了甘巴拉山乳胶室的实验结果. 在相互作用方面, HDSQI 模型考虑了 QCD 喷注, 假定 Feynman scaling 在碎裂区近似成立, 在中心区强烈破坏, 并且, 非弹相互作用截面随能量上升而显著增加. 对于原初成分, 该模型假定原初粒子分质子、氦核、轻重核、中重核、重核、甚重核和铁核, 在 10^{15} 到 10^{16}eV 能区附近以重的成分为主(详见文献[3]).

3.2 输入、输出参数的选取

本工作挑选描述族事例的特征量(共27个)作为ANN的输入参数，为了应用，这些量都可以通过实验测量得到．如下所示：

(1) 能量加权平均横向扩展〈ER〉和横向扩展〈R〉．这里的能量E是簇射能量，横向扩展R是族中簇射在观测面上与族能量中心的距离，对族中所有簇射的R和ER求平均，可得〈R〉和〈ER〉．

所选取的参数包括整个族事例的扩展〈ER〉、〈R〉，以及族中γ和强子的扩展$\langle ER\rangle_\gamma$、$\langle R\rangle_\gamma$和$\langle ER\rangle_h$、$\langle R\rangle_h$．这两种量与相互作用的次级粒子的横动量有关．

(2) 观测能和簇射数．输入参量选取族中强子的观测能ΣE_h和簇射数n_h，族中γ的观测能ΣE_γ和簇射数n_γ，族事例总的观测能和簇射数以及n_h / n_γ和$\Sigma E_h / \Sigma E_\gamma$，其中，最小簇射能为4TeV．

(3) 天顶角θ．θ大小直接影响族事例的几何形态．例如，对两个完全相同的族事例而言，天顶角大的，它的横向扩展也大．

(4) Q_{max}分布．Q_{max}表示族中能量最大簇射占总观测能的份额．该量主要用于描述相互作用中的领头粒子，带有较多原初成分的信息，并与非弹系数密切相关．

(5) 族事例核心的行为．初级粒子与族事例核心行为的关联在文献[11]中已有详尽论述．所采用的输入参量是：

$F_5=E / \Sigma E_\gamma(R<5\text{mm})$(族中横向扩展小于5mm的簇射的能量之和与族总观测能的比例)、n_5(族中横向扩展小于5mm的簇射的数目与族中簇射总数的比例)、F_{10}和n_{10}(横向扩展小于10mm时与上面相应的两个量)以及F_5 / F_{10}和n_5 / n_{10}．

(6) 集团效应．族事例中次级粒子集中于空间几个不同区域，形成集团．集团的特征关系到原始相互作用的性质，例如，多重产生中的大横动量硬散射过程可以产生集团现象．我们选择族中集团数目N_c和头三个能量最大的集团占族总观测能的份额F_{123}作为输入参量．N_c与相互作用多重数有关，而F_{123}对原初也有一定的敏感性．

(7) 族中高能量粒子的行为．定义$f'=E / \Sigma' E_\gamma$，其中，E为粒子簇射能，取阈值$f'_m=0.04$，则n'、$\Sigma' E_\gamma$分别表示满足条件$f' \geqslant f'_m$的粒子簇射数目以及簇射能总和．输入参数为$\Sigma' E_\gamma$和n'，以及两者的比例．这些量反映了族中高能簇射的行为，对初级成分有一定敏感性[3]．

将输出量(初级粒子)分成两类，即要分辨质子和原子核时，将质子单独算作一类，而将原子核算成另一类；要将铁核与其它成分区分开，则把铁核算作一类，其它成分算另一类．

3.3 GCCA的训练和测试

依据前面对输入、输出量的选取方法，使用GCCA进行训练和测试，样本中，观测能在100和500TeV间铅室族事例有1705个，由于观测能大于500TeV的事例流强非常低，相应的模拟机时很多，因此，这个能区的铅室族事例只有238个．训练集和测试集的样本数各取总样本数(即族事例总数)的50%．训练参数如下：群体大小为100，交

配率为0.6，突变率为0.001，最大隐节点数为50，终止准则为网络能正确区分全部训练样本或均方和误差下降到0.01并趋于稳定时，则停止训练．训练完成后，输入测试集样本以检验训练效果．

4 实验结果与讨论

网络的训练效果是以其学习精度，即训练后网络对训练集的正确响应率及其泛化能力，即训练后网络对测试集的正确响应率来衡量的，表1给出GCCA对训练集和测试集的判选结果．

由表1结果可见，用GCCA分辨质子和原子核，其正确率随着族总观测能的增加而上升，这与下述物理事实是吻合的，即能量很高的由质子产生的族事例的能流集中在族能量中心附近，而原子核产生的族则比较分散，能量越高，差别越明显，因此也越容易分辨．

为探讨网络对不同模型的敏感性，用由HDSQI模型训练出来的网络鉴别PFQI(火球)模型[3]的数据．该模型现有的数据是由质子产生的垂直入射族事例．总观测能大于500TeV的族有48个，训练集为238，测试集为48；总观测能在100和500TeV事例数共492个，训练集为1705，测试集为492，判选结果如表2所示．

表1 铅室事例的识别结果

族总观测能	100— 500(TeV)				>500(TeV)			
样本集	训练集		测试集		训练集		测试集	
输出	p	N	p	N	p	N	p	N
正确率	73.5%	72.3%	75.9%	70.3%	80.7%	84.8%	75%	87.3%
误判率	25.9%	28.3%	22.0%	32.2%	17.5%	16.4%	13.4%	23.8%

p和N分别表示质子和原子核．

表2 不同模型的识别结果

族总观测能	100— 500(TeV)		>500(TeV)	
样本集	训练集(HDSQI)	测试集(PFQI)	训练集(HDSQI)	测试集(PFQI)
输出	p	p	p	p
正确率	74.3%	61.6%	79.5%	79.2%
误判率	19.8%	38.4%	14.6%	20.8%

上述结果说明，在100到500TeV能区，测试与训练的结果相差较大，而在大于500TeV能区，两者差别较小，这可能预示着网络对不同模型的数据具有一定敏感性．以后，我们将用多种模型分不同能区专门研究这个问题．

我们还尝试了分辨铁核与其它成分．结果GCCA将样本中的大部分铁核判选成包括质子在内的其它成分．从物理上看，铁核与其它原子核所产生事例的各种特性是非常相近的，这也许是原因之一．另外，原初成分中，铁核在数目上远少于其它成分的总

和，比例约为 1 /10，在 100 和 500TeV 能区间的样本中总共才有 150 个，而在大与 500TeV 的能区中总共只有 24 个．这样的数据量对训练网络是远远不够的，因此，GCCA 对铁核的分辨能力目前还不能确定，下一步将用足够大的样本重新进行判选，以确定网络对铁核的判选能力．

不同的神经网络对原初的判选能力不同．我们采用了一种应用较为广泛的基于 BP 算法的多层前馈式网络[12]对相同样本进行了判选，网络中包含 27 个输入节点，一个隐藏层，20 个隐节点和一个输出节点．训练次数为 100，学习速率取 0.6，矩为 0.8．结果表明，该网络对原初的分辨能力是很差的，而且，在解决其它复杂问题上[7]也远不如 GCCA，因此，所用网络的性能越好，对原初的分辨能力有可能越高．

从上面结果可以看到，人工神经网络能比较有效地分辨超高能区原初宇宙线当中的质子和原子核．今后我们将增加模拟数据量，尤其是观测能大于 500TeV 族的数据量，重新进行上面的实验，并逐步完善这种方法．有确切结果后，再利用网络对甘巴拉实验组的实验数据进行识别．

本工作曾得到山东大学高能物理教研室赵炘同学的帮助，模拟数据是由日方合作组提供的，在此表示感谢．

参　考　文　献

[1] L. Lonnblad *et al.*, *Phys. Rev. Lett.*, **65** (1994) 1321.
[2] G. Stimpfl－Abele *et al.*, *Computer Phys. Commun.*, **64** (1991) 46.
[3] J. R. Ren *et al.*, *Phys. Rev.*, **D38** (1988) 1404.
[4] D Goldberg, Genetic Algorithmsin Search, Optimization and Machine Leaning, Massacheusettes: Addison－Wesley Pub. 1989.
[5] S. E. Fahlman, Tech Report., CMV－CS－90－100.
[6] Hualou Liang *et al.*, *IEEE* ICNNSP'95, p56 — 59, 1995.
[7] 梁化楼，中国科学院高能物理研究所博士学位论文，1996.
[8] B. Widrow *et al.*, 1960 IRE WESCON conr. Record, 96 — 104, Aug. 1960.
[9] S. E. Fahlman, Proc. of the 1988 Connectionist Model Summer Schools 1988, 38 — 51.
[10] C. M. G. Lattes *et al.*, *Phys. Rep.*, **65** (1980) 151.
[11] 任敬儒等，高能物理与核物理，**16** (1992) 193.
[12] D. E. Rumelhart *et al.*, Parallel Distributed Processing, Vol. 1 — 2, MIT Press, 1986.

Distinguishing Primary Cosmic-Ray Composition with Artificial Neural Networks

Liang Hualou Xie Wei Ren Jingru
Wang Taijie Dai Guiliang
(*Institute of High Energy Physics, The Chinese Academy of Science, Beijing* 100039)

Received 5 February 1996

Abstract

We used artificial neural networks (ANN) to distinguish superhigh energy cosmic-ray proton (p) and nucleus (N) with Monte Carlo family data in mountain emulsion chamber experiment. The result shows that when visible energy of a family is larger than 500TeV, about 80% of p and N can be correctly selected, and more than 70% can be selected in the region between 100 and 500TeV.

Key words neural networks, genetic algorithm, primary cosmic-ray composition, mountain emulsion chamber.