文章编号: 1007-4627(2013) 04-0446-05

基于 Bagging 算法的神经网络在粒子鉴别中的应用

崔向丽[1, 3]，陈旭荣[2, 3]，喻梅凌[4]，周庆国[1]

(1. 兰州大学信息科学与工程学院，甘肃 兰州 730000;
2. 中国科学院近代物理研究所，甘肃 兰州 730000;
3. 兰州强子物理和CSR物理研究中心，甘肃 兰州 730000;
4. 杭州电子科技大学信息工程学院，浙江 杭州 310018)

摘要: 分析了神经网络方法和bagging算法在实验高能物理和核物理数据分析中的应用现状。分别对神经网络方法和bagging算法的基本原理进行了介绍。以蒙特卡罗产生器产生的夸克胶子喷注样本为例，详细讨论了神经网络方法以及bagging算法与神经网络结合对粒子鉴别中信号和背景区分问题的应用过程，并对结果进行了讨论和分析。实验结果表明，应用bagging算法后，神经网络能够较大幅度地提高实验高能物理和核物理数据分析中粒子鉴别的精度，以及能够得到较高的信噪比。

关键词: 数据分析；bagging算法；粒子鉴别；神经网络

中图分类号: O572.21^{+}3　　**文献标志码:** A　　**DOI:** 10.11804/NuclPhysRev.30.04.446

1 引言

在实验高能物理和核物理数据分析中，区别信号与本底是需要解决的最普遍最重要的问题之一[1]。粒子鉴别问题实质是高能物理领域中的模式识别问题。近年来，人们发展了很多的用于解决此类问题的方法。例如，贝叶斯方法、决策树法、k近邻法和线性判别方法等。其中，应用最为广泛的是Fisher线性判别方法和神经网络方法。上述方法在预测能力、可解释性、稳定性、训练以及分类所消耗的时间上都有各自的优势，在实际应用中要根据分析的实际需要来选择合适的分类器。

在这些分类方法中，神经网络因其独特的并行计算和非线性处理能力在模式识别领域有着重要的应用。早在1988年，人工神经网络方法就被引入高能物理实验领域[2]并广泛地应用于夸克胶子喷注的鉴别[3]、电子强子分辨、顶夸克[4]和Higgs粒子的寻找[5]。同时，近年来其在核物理及核工程中也有较广的应用[6−7]。但是，在实际中，神经网络也存在性能不稳定的问题，即其对训练样本有相对较高的依赖性。此外，网络结构和输入初始权值等都会影响神经网络的判别能力。目前，在理论上还没有确定的方法来估计这些因素的影响，也没有具体的方法来指导参数的选择，神经网络的应用效果还取决于使用者的经验。对于神经网络和决策树这样的弱分类器，可以通过使用网络集成的方法，来提高神经网络的判别能力。本文将讨论运用bagging算法如何提高神经网络对高能物理中的夸克喷注和胶子喷注的判别能力。

2 方法简述

2.1 神经网络方法概述

神经网络按照其网络拓扑结构及运行方式，可以分为前馈网络、反馈网络和横向网络。人工神经网络优化权值的方法称为学习，学习规则就是修改网络的权值和偏置值的方法和过程。学习规则可以分为有监督学习、无监督学习以及增强学习。目前，在模式识别中应用较成熟的神经网络模型是误差反向传递(Back-Propagation)算法的前馈网络(简称BP网络)。BP网络不仅有输入和输出节点，而且还有一层或多

收稿日期: 2012-12-27;　**修改日期:** 2013-01-15

基金项目: 国家自然科学基金资助项目(11005045, 11175220)；中国科学院百人计划资助项目(Y101020BR0)

作者简介: 崔向丽(1987–)，女，山西临汾人，硕士研究生，从事神经网络研究；E-mail: cuixl10@lzu.edu.cn.

通信作者: 陈旭荣，E-mail：xchen@impcas.ac.cn

http://www.npr.ac.cn

层隐含节点，对于输入信息来说，先要向前传播到隐层节点上，经过隐含节点的输出信息传播到输出节点，最后给出输出结果。网络的学习过程由正向和反向传播两部分组成。在正向传播过程中，每一神经元的状态只影响到下一层神经元的状态。如果网络的实际输出与期望输出之间的误差超过可接受范围，那么转入反向传播过程，即将误差信号沿原来的连接通路返回，并逐层地修改各层神经元的连接权值，修改完成后，再将训练样本经过网络正向传播，重复以上过程，直到误差信号达到可接受的范围。训练过程结束后，网络记录了最后更新的参数值，BP 网络就可以用于识别新的样本。

2.2　bagging方法

bagging (Bootstrap aggregating) 是 Breiman[8] 于 1996 年正式提出[9]，其过程是给定一弱学习算法和样本集，每次从训练样本中取样 m 个 ($m < n$，n 为样本总体) 进行训练，训练完毕后将其放回样本中，再进行下一轮的取样训练。初始训练序列可能在子训练集中出现多次，也可能根本没有出现，训练完成后生成一个预测函数序列，最终的预测结果将由这些子训练函数按照一定的集成方法来得到。

在集成网络的个体方法找到后，集成网络要对训练后的结果进行集成，集成的方法通常有以下两种。

(1) 平均加权方法：采用这种方法，集成的输出直接由各子网络的输出做简单的平均或加权平均产生。

(2) 非线性结合方法：非线性结合方法包括顺序统计方法、投票法等。顺序统计法是将子网络的输出结果标记为 1 或者 -1，然后顺序将子网络的输出相加，得到最终结果为大于 0 或者小于 0 来判别类别。投票法通常采用绝对多数的投票作为结果，即当且仅当有超过半数的输出结果为某种分类时，最终的结果才被确定为该分类。有时也使用相对多数投票的方法，即当且仅当某输出结果取得数目最大的投票时，网络的最终结果将被确定为对应的类别。理论分析和大量的实验表明，相对多数投票法要优于绝对多数投票方法，因此，多数分类器在集成时多采用相对多数投票法。本实验中采用了相对多数投票的方法对结果进行集成。

2.3　bagging方法与神经网络的结合

本实验中将神经网络方法与 bagging 方法结合，用 bagging 取样方法对每一个子神经网络进行取样，训练后放回，再对下个子神经网络进行同样的操作。假设设有 N 个样本，要训练的神经网络总数为 T，i 为被训练的第 i 个子神经网络的标号。具体的实现过程如下：(1) 初始化样本。设 $i = 1$；(2) 从初始样本中随机抽取 M 个样本 ($M < N$)，训练第 i 个神经网络，训练完成后，$i = i + 1$；(3) 若 $i \leqslant T$，返回 (2)，否则，转 (4)；(4) 用训练好的 T 个神经网络进行集成，集成的方法如下：

$$f_{\text{final}}(x^p) = \begin{cases} 1\text{，if} & \sum\limits_{t=1}^{T} f_t(x^p) \geqslant \dfrac{1}{2}T \\ 0\text{，if} & \sum\limits_{t=1}^{T} f_t(x^p) < \dfrac{1}{2}T\text{。} \end{cases} \tag{1}$$

3　实验与结果

3.1　样本的选择

本实验的数据来自于文献 [10]。样本是由蒙特卡罗事件产生器 Jetset 7.4[4] 产生的质心能量为 91.2 GeV 的正负电子碰撞事件。在这一碰撞能量下正负电子湮灭产生一个高能虚光子或 Z 玻色子，然后这些中间粒子衰变成两个正反夸克。当正反夸克能量较高时，有可能辐射出一个具有较高横能量的胶子。夸克和胶子在末态碎裂为强子，从而在探测器中形成末态三喷注事件。数据分析中需要解决的问题是，确定末态强子喷注来源于夸克还是胶子。在试验中无法直接记录夸克和胶子 (部分子) 如何强子化到强子，但在蒙特卡罗模拟中可以将部分子层次和强子层次对比，通过部分子信息得到强子喷注的类别，从而得到训练神经网络的夸克喷注和胶子喷注样本。具体做法如下。

(1) 对每一事件，分别在部分子层次和强子层次利用 Durham 喷注判定法[11]将事件中的粒子分成 3 个喷注，强子层次选用的粒子为带电强子。

(2) 选择 3 个喷注落在同一个平面上的事件，即要求在强子层次三喷注内相邻喷注之间的夹角满足条件 $\theta_1 + \theta_2 + \theta_3 > 358°$，而该条件在部分子层次自动满足。

(3) 在强子三喷注中，应用角度截断法[12]挑选出夸克喷注和胶子喷注，做法为：将相邻喷注间的夹角排序，定义与最大夹角相对的喷注为胶子喷注，与最小夹角相对的喷注为能量较大的夸克喷注。本实验中，要求最大夹角与次大夹角之间的差别大于一个角

http://www.npr.ac.cn

度截断。

(4) 在部分子三喷注中，通过计算喷注内部分子的净夸克数可以知道喷注是源于夸克还是胶子。在这一步骤中，运用了动量守恒定律，通过检查强子喷注动量相对部分子喷注动量的角度偏移来确认该强子喷注是否源于该部分子喷注。最后去掉最大能量的夸克喷注，而保留次大能量的夸克喷注和胶子喷注为我们的样本。

(5) 本实验中限定了能量为 2.5 ~ 22.5 GeV 的喷注为最后所采用的喷注样本。

本实验数据记录了喷注的 7 个特征参量，分别为喷注多重数 n、喷注内粒子平均横动量 P_t、喷注所对的夹角 θ、喷注能量 ε、喷注内粒子的总横动量 $P_{t,\,total}$、喷注内粒子相对喷注轴的平均夹角 θ_{avg} 以及喷注内粒子相对喷注轴的夹角之和 θ_{avsum}。在夸克胶子喷注的识别中，选择反映两者不同性质的特征物理量至关重要。量子色动力学(QCD)理论预言在相同喷注能量下胶子喷注的多重数比夸克喷注大且横动量分布比夸克喷注宽[13−15]，所以本实验中选择至少4个物理量作为特征量：n，P_t，θ 和 ε 作为神经网络的输入。为了验证结果的有效性，实验中也尝试了选择其他的物理量作为神经网络的输入特征量。

3.2 神经网络的设置

本实验选取 2 500 个夸克喷注 (信号，设期望输出为 1) 和 2 500 个胶子喷注 (背景，设期望输出为 −1) 作为训练样本。另取同样比例的 5 000 个喷注作为测试样本。定义一些参量来描述分类器的分类效率。假定测试样本中有 N_s 个信号和 N_b 个背景，分类器应用于数据集后得到的结果为：信号中被判定为信号的为 N_{cs}，被误判的为 N_{ws}，背景中被正确判定为背景的有 N_{cb} 个，被误判的有 N_{wb} 个。用粒子纯度、分类效率以及信噪比来描述分类器的性能，定义式如下：

$$purity = \frac{N_{cs}}{N_s}\ ,\quad errorratio = \frac{N_{ws}+N_{wb}}{N_s+N_b}\ ,\quad signal_bg_ratio = \frac{N_{cs}}{N_b-N_{cb}}\ 。\tag{2}$$

在本实验中，选用的神经网络为 BP 网络[16]，训练方法选择最速下降法[17−18]，当网络对于训练样本的误差小于 0.005 时，停止训练，并保存最新的权值和阈值。在表 1 中，用 X-X-X 来分别表示不同数量的输入层、隐含层和输出层。为辨别神经网络的输出，定义一个接收度 σ，当输出层的神经元个数为 1 时，规定输出层输出 output 大于 $(1-\sigma)$ 时为夸克 (信号)，而 output 小于 σ 时为胶子 (背景)；当输出层的神经元个数为 2 时，以 (1, 0) 的序对来表示夸克 (信号)，以 (0, 1) 的序对来表示胶子 (背景)，类似的神经网络的两个输出 output1、output2 以上述类似的方式来区分夸克 (信号) 与胶子 (背景)。表 1 所列为不同的网络结构，单个神经网络对粒子的分类效率以及运用 bagging 算法后分类器的分类效率。

表 1 运用神经网络方法以及 bagging 算法与神经网络结合后对粒子鉴别的结果

算法	网络结构	夸克纯度	胶子纯度	信噪比	误判率
应用神经网络方法	4-5-1	0.842	0.934	12.675	0.112
	4-5-2	0.923	0.947	17.353	0.065
	5-4-1	0.874	0.891	8.007	0.117
	5-4-2	0.8556	0.961	22.0515	0.092
应用 Bagging 算法	4-5-1	0.986	0.980	50.302	0.017
	4-5-2	0.985	0.982	54.298	0.016
	5-4-1	0.986	0.981	51.012	0.016
	5-4-2	0.987	0.979	47.294	0.017

从表中可以得出下述结论，传统神经网络能够有效地用于夸克和胶子的鉴别，鉴别纯度已经可达到 80% 及其以上，可见，神经网络在粒子鉴别中已有其独特的优势。另外，当所选择的网络有 2 输出元时，得到的纯度要优于输出神经元为 1 的网络，并且有较高的信噪比以及较小的误判率。在应用 bagging 算法

http://www.npr.ac.cn

之后，神经网络对测试样本的纯度更有大幅度的提高，这主要是因为 bagging 算法是基于一种投票的机制，单个神经网络可能对样本具有一定的误判，但是如果采用多个神经网络的投票，则可以大大降低误判率，提高神经网络的预测性能。同时，bagging 算法对于提高信噪比以及降低误判率也都有明显的作用。

为了印证以上结论的正确性，我们选择了不同的网络结构进行测试，对比的结果均与上述有明显的相似性。其中网络结构为 4-5-2 时，选用的网络的输入为喷注多重数 n，喷注内粒子总横动量 $P_{\mathrm{t,\ total}}$、喷注所对的夹角 θ、喷注能量 ε。选用同样的网络结构，但将第2个输入参量换为喷注内粒子平均横动量，得到的效率和纯度将提高 1% ～ 7%，这说明平均横动量比总横动量更能描述夸克喷注和胶子喷注的不同性质。同时在本实验中，也选择了其他的属性作为神经网络的输入，结果显示当神经网络不含有输入 n，P_{t} 以及 θ 时，网络会陷入混乱，具体表现为网络有极大的误判率以及较低的识别纯度。而 $P_{\mathrm{t,\ total}}$、θ_{avg} 和 θ_{avsum} 的引用与否对实验的结果都不会有太大的影响，但增加了网络的训练时间以及检索时间。因此为了减少神经网络的运算时间以及存储空间，实际中可以考虑不将其作为神经网络的输入。

4　总结与展望

本文主要描述了用 bagging 算法来提高神经网络对实验高能物理和核物理数据的鉴别能力，比较了不同网络结构的判别能力，讨论了使用不同输入参量的平衡作用的重要性和方法。由此，可以得出一些关于应用神经网络的普遍的规律，即对于具体问题网络结构不一定越复杂越好，在实践中可以尝试先从简单的开始；输入参量不是越多越好，需要在试验中找出对解决问题最关键的因素，平衡网络的训练时间，以及网络的实用性能之间的关系。

通过本实验，还可以得出，神经网络在高能物理数据处理方面已有一定的优势。例如，有较高的粒子鉴别纯度和较高的信噪比等。在引入 bagging 算法后，分类器的性能在各方面都得到了很大的提升。在后续的研究中，我们将尝试将其应用于不同的数据集，去测试方法的通用性和实用性。同时将探索 bagging 算法在实验高能物理核物理数据分析中更为广泛的应用。

参考文献(References)：

[1] ZHU Yongsheng. Experimental Data Multivariate Statistical Analysis[M]. Beijing: Science Press, 2009: 88–134. (in Chinese)
(朱永生. 实验数据多元统计分析[M], 北京: 科学出版社, 2009: 88–134.)

[2] DENBY B. Computer Physics Communications, 1988, **49**(3): 429.

[3] LÖNNBLAD L, PETERSON C, RÖGNVALDSSON T. Physical Review Letter, 1990, **65**: 1321.

[4] SJÖTRAND T, EDÉN P, FRIBERG C, *et al*. Computer Physics Communications, 2001, **135**: 238.

[5] MAGGIPINTO T, NARDULLI G, DUSINI S, *et al*. Physics Letters B, 1997, **409**: 517.

[6] YU Guoliang, HOU Long, ZHANG Xiaopeng, *et al*. Nuclear Physics Review, 2012, **29**(3): 272. (in Chinese)
(于国梁, 侯龙, 张笑鹏, 等. 原子核物理评论, 2012, **29**(3): 272.)

[7] XIE Junhua, LIU Zhigui, REN Lixue, *et al*. Nuclear Physics Review, 2012, **29**(2): 202. (in Chinese)
(谢军华, 刘知贵, 任立学, 等. 原子核物理评论, 2012, **29**(2): 202.)

[8] BREIMAN L. Machine Leaning, 1996, **24**(2): 123.

[9] GUO Zhaobin. Noise Tolerating Capability of Bagging-based Neural Networks and Their Application[D]. Dalian: Dalian University of Technology, 2009: 4–10. (in Chinese)
(郭照斌. Bagging 算法神经网络的抗噪声能力及应用[D]. 大连: 大连理工大学, 2009: 4–10.)

[10] YU Meiling, ZHANG Kunshi, LIU Lianshou. International Journal of Modern Physics A, 2005, **20**: 7603.

[11] DOKSHITZER Y L, KHOZE V A, TROYAN S I. Physics Letters B, 1992, **274**: 116.

[12] YU Meiling, LIU Lianshou. Chinese Physics Letters, 2002, **19**: 647.

[13] MUELLER A H, QIU Jianwei. Nucl Phys B, 1986, **268**: 427.

[14] GAFFNEY J B, MUELLER A H. Nucl Phys B, 1985, **250**: 109.

[15] MALAZA E D, WEBBER B R. Phys Lett B, 1984, **267**: 702.

[16] HAYKIN S. Neural Networks: A Comprehensive Foundation[M]. YE Shiwei(2nd ed), Beijing: China Machine Press, 2004: 17–123. (in Chinese)
(海金. 神经网络原理[M]. (第二版). 叶世伟, 译, 北京: 机械工业出版社, 2004: 17–123.)

[17] HAGAN M T. BEALE Mark Hudson. Neural Network Design[M]. DEMUTH H B, Beijing: China Machine Press, 2002: 40–89. (in Chinese)
(哈根. 神经网络设计[M]. 戴葵, 译, 北京: 机械工业出版社, 2002: 40–89.)

[18] RIPLEY B D. Pattern Recognition and Neural Networks[M]. Beijing, Posts & Telecom Press, 2009: 73–107. (in Chinese)
(里普利. 模式识别与神经网络[M]. 北京: 人民邮电出版社, 2009: 73–107.)

http://www.npr.ac.cn

Application of Bagging Algorithm Based on Neural Network in Particle Identification in Data Analysis

CUI Xiangli[1]，CHEN Xurong[2, 3]，YU Meiling[4]，ZHOU Qingguo[1]

(*1. School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China;*
2. Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou 730000, China;
3. Lanzhou hadron physics and CSR Physics Research Center, Lanzhou 730000, China;
4. Hangzhou Dianzi University Information Engineering School, Hangzhou 310018, China)

Abstract: The paper presents the application of neural network and bagging algorithm in experimental high-energy physics and nuclear physics data analysis. Paper also introduces the basic principles of neural network method and bagging algorithm. We use the data samples of quark-gluon jets, which are generated by Monte Carlo generator, to solve the problem of discriminating signal events and background events by the combined algorithm of bagging algorithm and neural network. Experimental results show that, to apply bagging algorithm, neural networks can greatly improve the accuracy of the identification of particles in the experiments of high energy physics and nuclear physical data analysis, and also obtains a larger SNR (Signal to Noise Ratio).

Key words: data analysis; bagging algorithm; particle identification; neural network

Received date: 27 Dec. 2012； **Revised date:** 15 Jan. 2013

Foundation item: National Natural Science Foundation of China (11005045, 11175220); Hundred Talents Program of Chinese Academy of Sciences(Y101020BR0)

Corresponding author: CHEN Xurong, E-mail: xchen@impcas.ac.cn

http://www.npr.ac.cn