机器学习在宇宙线粒子鉴别中的应用*

刘烨[1)] 牛赫然[1)] 李兵兵[2)] 马欣华[3)4)] 崔树旺[2)†]

1) (河北经贸大学管理科学与工程学院, 石家庄 050061)

2) (河北师范大学物理学院, 石家庄 050024)

3) (中国科学院高能物理研究所粒子天体物理重点实验室, 北京 100049)

4) (四川天府新区宇宙线研究中心, 成都 610000)

(2023 年 3 月 7 日收到; 2023 年 4 月 8 日收到修改稿)

基于热中子探测器实验模拟数据, 使用决策树 (decision tree, DT)、随机森林 (random forest, RF) 和 BP 神经网络 (back-propagation neural network, BPNN) 构建了宇宙线粒子鉴别机器学习模型, 对每种粒子分别使用不同的机器学习算法基于模拟数据进行模型训练, 并针对算法进行超参数调整, 将每种算法的 AUC 值和 Q 品质因子作为粒子成分鉴别的评价指标. 实验结果表明, 不同机器学习模型对粒子预测精度影响很大. 在测试检验中, 经过交叉网格搜索方法调参后的决策树鉴别模型对中成分 (碳氮氧和镁铝硅) 比较敏感, 鉴别模型 AUC 值均在 0.95 以上, Q 品质因子均大于 6; 经交叉网格搜索方法调参后的随机森林鉴别模型对于宇宙线粒子鉴别的效果最好, 所有粒子鉴别模型的 AUC 值均大于 0.92 且 Q 品质因子均在 4 以上; BP 神经网络算法只对质子和铁核比较敏感. 本研究对宇宙线粒子鉴别和筛选提供了新的方法和选择, 可为热中子探测器后续开展宇宙线能谱测量提供新思路.

关键词: 宇宙线, 粒子鉴别, 机器学习, 随机森林

PACS: 02.70.–c, 95.55.Vj **DOI:** 10.7498/aps.72.20230334

1 引 言

宇宙线是唯一来自外太空的物质样本, 本质是高能带电粒子流, 能量从 keV 到 EeV 跨越 17 个量级, 并且在传播过程中会与星际物质相互作用产生少量次级核子和反质子、反电子、伽马光子、中微子等次级宇宙线粒子[1–3]. 在宇宙线研究领域中, 宇宙线能谱结构和次级宇宙线粒子成分的精确测量是解决宇宙线起源、加速、传播机制等问题的关键[4,5]. 目前, 多个实验已经测量到了宇宙线能谱中的“膝区”结构, 但是“膝区”的确切位置及成分存在较大差异[6], 因此精确鉴别宇宙线中的粒子成分十分重要, 是开展相关科学研究的重要基础和前提.

传统宇宙线成分鉴别大多基于多变量分析方法完成, 该方法需要人工选取特征, 耗费人力资源的同时容易丢失数据信息[7], 而机器学习方法能直接在原始数据的基础上进行分析, 节省人力资源的同时尽可能挖掘数据的信息. 机器学习是人工智能的分支之一, 是统计学、人工智能和计算机科学交叉的研究领域, 可以通过学习多源、复杂的数据内在模式和结构, 挖掘隐藏在数据背后的信息, 并用于解决分类、回归、聚类等复杂问题[8]. 随着机器学习的不断完善和计算能力的提升, 机器学习算法也逐渐帮助科研人员分析和处理大量的物理学相关数据. Herrera 等[9] 评估了人工神经网络 (ANN)、

* 国家自然科学基金 (批准号: 11905043, U2031103)、河北省教育厅项目 (批准号: KCJSZ2022036) 和河北经贸大学研究生创新资助项目 (批准号: XYCX202333) 资助的课题.

† 通信作者. E-mail: cuisw@hebtu.edu.cn

极端梯度提升树 (XGBoost)、支持向量机 (SVM) 和 K 近邻 (KNN) 算法对超高能宇宙线成分的分类效果, 并使用五折交叉验证的方法对算法的超参数进行优化, 结果表明极端梯度提升树对所有成分都表现出优异性能, 准确率和 f1 评分均为 0.97, 且运行时间最短, 支持向量机的准确率和 f1 评分均为 0.94, 但是运行时间较长, 人工神经网络和 K 近邻算法效果稍差; Pang 等[10] 在高能核物理领域利用卷积神经网络 (CNN) 模型, 将不同状态方程下相对论流体力学演化末态的粒子分布作为神经网络输入, 将演化使用的和物质状态方程种类作为标签做监督学习, 将寻找 QCD 相变临界点的任务转化为两个相变区域分类问题; 高泽鹏等[11] 使用 LightGBM 决策树算法训练初始化过程中有无形变效应给出的反应末态的自由质子、带点碎片及 π^+, π^- 的 $p_t - y_0$ 谱, 通过碰撞末态数据反推初态结构, 分类的准确率在 60%—70% 之间, 同时, 此研究还通过 LightGBM 决策树算法计算了特征重要性, 发现弹靶快度区形变的带电碎片敏感于弹靶核的初始形变, 与相关理论分析相一致.

本研究以热中子在探测器模拟数据为研究对象, 以粒子的原初能量、天顶角、电子数、中子数及芯距 5 个量作为特征, 应用决策树 (decision tree, DT)、随机森林 (random forest, RF) 和 BP 神经网络 (back-propagation neural network, BPNN) 3 种机器学习算法, 构建了 3 种宇宙线粒子鉴别模型, 并调整 3 种算法的超参数以提高其对宇宙线成分鉴别能力, 然后使用相关评价指标对这 3 种模型的结果进行评估, 得到了性能最优的鉴别模型. 最后, 用验证数据验证了最优鉴别模型的精度和泛化能力, 为后续开展宇宙线能谱精确测量提供依据和参考.

2 研究方法

本文选择决策树、随机森林和 BP 神经网络 3 种常用的机器学习算法建立宇宙线粒子鉴别模型. 实验中, 首先通过宇宙线粒子在探测器上的坐标计算出粒子的芯距, 并选择宇宙线粒子原初能量 (E_0)、天顶角 (theta)、中子数 (neutron_total)、电子数 (MIPs_total) 和芯距 (core_distance), 5 个量作为成分敏感特征值, 然后将 5 种成分的数据混合在一起, 定义模型输出值若为“0”则对应目标成分, 若为“1”则对应其他成分, 并将数据按 4:1:5 的比例随机的划分为训练集、测试集和验证集, 分别用于模型的训练、测试和泛化能力的检验, 并且在训练过程中根据模型和粒子成分鉴别的评价指标, 不断的对模型的超参数进行调整, 筛选出最优鉴别模型. 本文中机器学习模型的训练、测试和验证均基于 Python 语言中 scikit-learn 和 Pytorch 库实现, 技术路线图如图 1 所示.

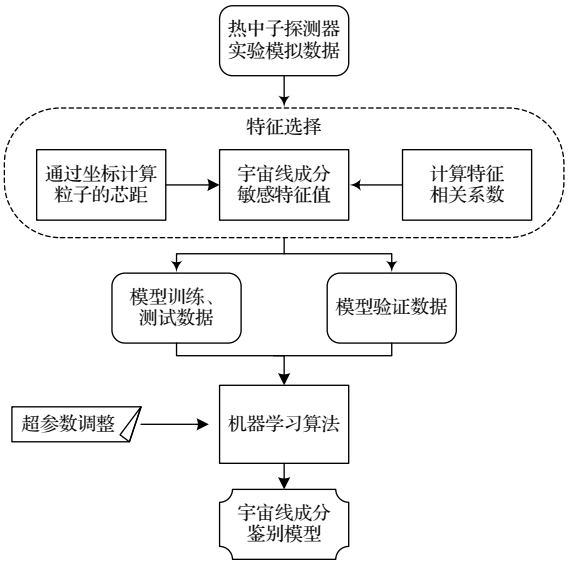

图 1 宇宙线成分鉴别模型技术路线图

Fig. 1. Technical roadmap of the cosmic rays component identification model.

为评估各机器学习鉴别模型对数据集分类的效果, 本文使用算法 AUC 值和宇宙线研究领域中的 Q 品质因子作为检验算法分类效果的评价指标. AUC 值等于 ROC 曲线下方面积, 是机器学习中一个通用的评价算法性能的指标, 用于权衡正确分类的收益和错误分类的代价之间的关联[12]. ROC 曲线分别以假正率 (FPR) 和真正率 (TPR) 为 x 轴和 y 轴:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}, \tag{2}$$

其中, TP 表示真正类, 即被模型预测为正类的正样本数; FP 为假正类, 即被模型预测为正类的负样本数; TN 为真负类, 即被模型预测为负类的负样本数; FN 为假负类, 即被模型预测为负类的正样本数.

热中子探测器模拟数据鉴别是一个分类问题, 但不能只使用统计学中常用的准确率判别模型分

类好坏, 因此本文使用高能物理领域中一个常用的评价指标 Q 品质因子对模型区分效果进行衡量[7], 其定义为

$$Q=\frac{\mathrm{Per_p}}{\sqrt{\mathrm{Per_e}}}, \tag{3}$$

其中 $\mathrm{Per_p}$ 为挑选目标成分的保留率, $\mathrm{Per_e}$ 为宇宙线其他成分的保留率.

2.1 数据集建立及预处理

本文使用的热中子探测器模拟数据由 CORSIKA 软件模拟生成, 该软件包含多种粒子反映模型, 可以模拟粒子到达不同海拔高度的相关信息, 包括粒子种类、能量、天顶角等, 这些参数已经得到了实验证实, 应用在众多宇宙线相关领域的实验中[13]. 热中子探测器模拟分为两部分, 首先利用 CORSIKA 软件模拟宇宙线在大气中级联簇射过程, 产生宇宙线粒子原初能量、天顶角、方位角及粒子位置等信息, 然后利用 Geant4 工具包开展热中子探测器响模拟. 最终热中子探测器模拟数据为质子、氦核、铁核、镁铝硅、碳氮氧, 每种成分各 4000 个事例, 能量范围为 1—10 PeV, 天顶角 0°—60°, 方位角为 0°—360°.

冗余特征可能会造成模型效率低或者过拟合等问题[14], 因此本文在构建特征过程中首先根据粒子位置信息计算出粒子到探测器中心的芯距, 并用其代替粒子其他位置信息, 作为特征加入到模型训练和测试过程. 因此, 本文在建模过程中使用宇宙线粒子的原初能量、天顶角、电子数、中子数及芯距 5 个量作为特征.

2.2 机器学习模型构建

2.2.1 决策树模型构建

决策树算法 (DT) 是一种经典的机器学习算法, 因其结构简单、学习成本低且可解释性强, 在机器学习领域有着广泛应用, 常用的决策树算法有 ID3, C4.5, CART 算法等[15]. 决策树的构建过程就是根据数据的不同特征, 将数据划分到不同区域, 使得同一区域的数据尽可能是同一种类型. 决策树算法构建过程是选择具有较强分类能力的特征生成决策树, ID3 算法是采用信息增益作为选择选择特征的度量, 而 C4.5 算法采用信息增益比[16]. 但由于决策树算法具有强大的建模能力, 因此会产生过拟合的问题, CART 算法在特征选择时以基尼系数为度量, 然后对所有属性可能进行遍历, 选择划分子集后基尼系数最小的节点进行分支, 这样可以简化树的结构, 避免过拟合问题[17]. 在信息论中, 信息熵用于描述变量分布的不确定性, 决策树在划分子树时以信息熵为基础, 进行相关计算, 然后选择特征划分子树. 对于离散型随机变量 D, 其信息熵为

$$H(D)=-\sum_{k=1}^{K}\frac{|D_k|}{|D|}\log_2\frac{|D_k|}{|D|}, \tag{4}$$

式中, K 为样本类别总数, $|D_k|$ 为第 k 类样本的数目, $|D|$ 为数据集 D 的数目. 使用特征 A 对变量 D 的条件熵为

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i), \tag{5}$$

则选择 A 构建子树的信息增益、信息增益比和基尼系数分别为

$$\begin{aligned}
&g(D,A)=H(D)-H(D|A),\\
&g_r(D,A)=\frac{g(D,A)}{H_A(D)}=\frac{g(D,A)}{-\sum\limits_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}},\\
&\mathrm{gini}(D)=1-\sum_{k=1}^{K}\left(\frac{|D_k|}{|D|}\right)^2.
\end{aligned} \tag{6}$$

本文建模过程中, 使用交叉网格搜索方法, 对树的深度最小分割样本数和最小分割叶子节点数等主要超参数进行调整. 交叉网格搜索方法是指定超参数取值的一种穷举搜索方法, 用于搜索算法的最优超参数组合. 通过将需优化算法的超参数运用交叉验证的方法进行优化, 即将各个超参数可能的取值进行排列组合, 列出所有可能的组合结果生成“网格”, 然后将各组合用于算法训练, 并使用交叉验证的方法对表现进行评估, 将平均得分最高的超参数组合作为最佳的选择, 返回给算法[18]. 决策树算法使用交叉网格搜索方法进行调整超参数时, 将表 1 所示的超参数设置在指定范围内, 将参数 cv 设置为 4, 其他参数默认, 搜寻最佳超参数组合. 决策树算法鉴别各种成分最佳超参数如表 1 所示.

表 1 决策树鉴别不同成分最佳超参数

Table 1. Optimal hyperparameters of decision tree identifying different components.

超参数	目标成分				
	质子	氦核	碳氮氧	镁铝硅	铁核
criterion	Entropy	Entropy	Entropy	Entropy	Entropy
max_depth	21	29	40	28	19
min_samples_split	2	4	7	2	4
min_weight_fraction_leaf	0	0	0	0	0
min_samples_leaf	1	1	1	1	1

2.2.2 随机森林模型构建

随机森林算法 (RF) 是一种监督机器学习算法, 广泛用于解决分类和回归问题. 本质上, 其是由多个决策树集成之后构建的, 使用 Bagging (自助聚类) 方法训练而成, 通过随机有放回的抽样方式选取数据构建分类器, 最后 通过组合学习得到的算法提升算法整体效果 [19]. 随机森林结构如图 2 所示.

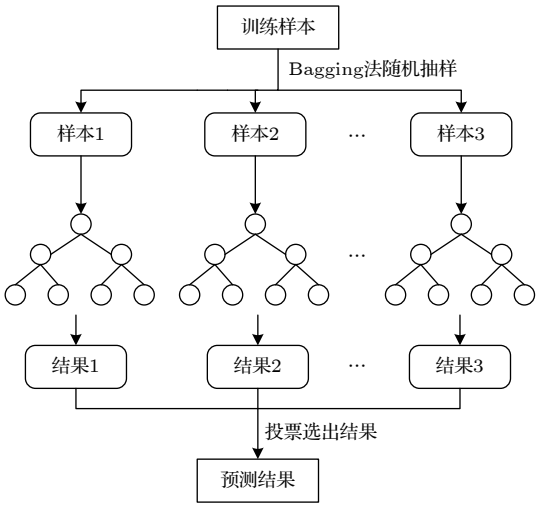

图 2 随机森林算法建模流程图

Fig. 2. Flow chart of random forest algorithm modeling.

随机森林算法可以看作是对原有决策树算法的整合和改进, 能够很好地处理变量间的非线性关系, 有着分类准确率高、抗噪能力优异、抗过拟合能力较强以及能够平衡非平衡数据的误差等优点; 此外, 随机森林算法能够在观测变量较少的前提下完成分类任务, 适合宇宙线粒子这种非平衡数据的分类 [20]. 本文使用随机森林算法建立宇宙线粒子成分鉴别模型过程中, 使用交叉网格搜索方法进行算法超参数调整, 调整结果如表 2 所示.

2.2.3 BP 神经网络模型构建

人工神经网络算法 (ANN) 是一种常用的非线性数据建模算法, 通过学习寻找并建立输入数据和目标数据之间的映射关系, 十分适合解决非线性和不确定性问题. BP 神经网络, 即前馈神经网络是一种多层前馈的人工神经网络, 其基本原理是输入信号前向传播, 误差反向传播 [21]. 在前向传播过程中, 输入信号经过输入层和隐藏层处理后, 到达输出层后输出. 若输出结果与预期结果不一致, 则根据预测误差, 使用梯度下降算法 (gradient descent) 调整各层网络的权重和偏置, 使得算法输出结果无限逼近预期结果, 直至得到损失不再降低或达到指定循环次数, 该过程称为反向传播 [22]. BP 神经网络结构一般分为 3 层, 即输入层、隐藏层和输出层, 输入层负责接收输入数据并转换为信号, 输出层负责输出模型结果, 隐藏层负责建立二者的映射关系. 本文 BP 神经网络结构示意图如图 3 所示.

表 2 随机森林鉴别不同成分最佳超参数

Table 2. Optimal hyperparameters of random forest identifying different components.

超参数	目标成分				
	质子	氦核	碳氮氧	镁铝硅	铁核
criterion	Gini	Gini	Entropy	Entropy	Entropy
n_estimators	48	88	30	15	21
max_depth	20	26	30	27	23
min_samples_split	2	2	2	1	2
min_samples_leaf	1	1	1	1	1

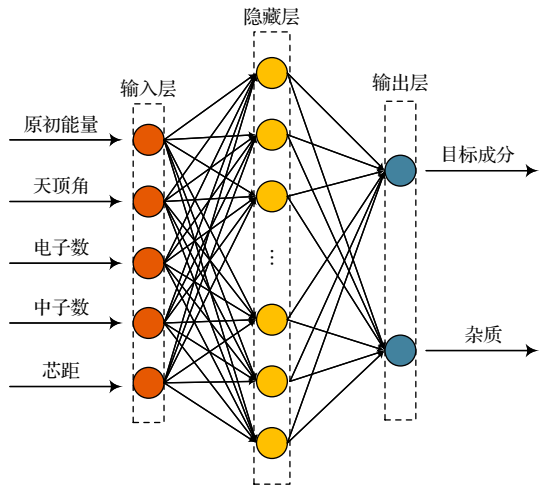

图 3 本文 BP 神经网络结构示意图

Fig. 3. Structure diagram of BP neural network in this paper.

隐藏层第 j 个神经元的输出值为 O_j, 计算公式为

$$O_j = \varphi(n_j) = \varphi\left(\sum_{i=1}^{N} \alpha_{ij} x_i + \lambda_j\right). \tag{7}$$

输出层第 k 个神经元的输出值为 O_k, 计算公式为

$$O_k = \psi(n_k) = \psi\left(\sum_{j=1}^{M} \beta_{jk} y_i + \gamma_k\right), \tag{8}$$

其中, n_j 和 n_k 分别为隐藏层第 j 个神经元和输出层第 k 个神经元的输入; α_{ij} 和 λ_j 分别为输入层第 i 个神经元到隐藏层第 j 个神经元的权重和偏置; β_{jk} 和 γ_k 分别为输入层第 j 个神经元到隐藏层第 k 个神经元的权重和偏置; N 和 M 分别代表输入层和隐藏层的神经元个数; ϕ 和 ψ 分别代表隐藏层和输出层的激活函数.

本文使用 BP 神经网络进行建模过程中, 首先对数据进行预处理以消除极端数据对于模型训练的影响, 数据预处理原理为

$$x_{\text{scalered}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{9}$$

其中 x_{scalered} 为标准化后的数据, $x_{\max}$ 和 $x_{\min}$ 分别为数据的最大值和最小值.

然后, 确定 BP 神经网络的拓扑结构. 本文中神经网络的输入和输出层均设置为一层, 输入层和输出层神经元个数分别设置为 5 个和 2 个, 隐藏层节神经元数由 Kolmogorov 公式 $N_{\text{h}} = \sqrt{N_{\text{in}} + N_{\text{out}}} + a$ 计算得出 [23], 其中 N_{h} 为隐藏层神经元数, N_{in} 为输入层神经元数, N_{out} 为输出层神经元数, a 为取值范围为 1—10 的常数. 实验中选取宇宙线粒子 5 个特征敏感值输入网络, 故 N_{in} 为 5; 实验中在输出层中通过 Softmax 函数计算并输入数据标签为 "0" 和 "1" 的概率, 故 N_{out} 为 2. 因此隐藏层节点数的取值范围是 $N_{\text{h}} \in [3, 13]$. 然后, 为了确定最佳隐藏层节点数, 采用控制变量法, 使用动态调整学习率算法, 初始学习率设置为 0.01, 每迭代 2000 次, 学习率变为原来的 0.7 倍, 其余条件不变, 只改变隐藏层节点个数, 并通过损失函数图像确定迭代次数, 进行模拟实验. 以鉴别氦核为例, 采用 BP 神经网络算法核验结果如表 3 所示.

综合考虑 AUC 值和 Q 品质因子, 确定隐藏层节点数为 13, 因此本文使用的 BP 神经网络结构为 5-13-2 的拓扑结构, 对热中子探测器中的氦核模拟数据进行鉴别. 表 3 给出本文根据评价指标确定 BP 神经网络算法鉴别氦核最佳拓扑结构的核验结果, BP 神经网络鉴别其他成分最佳超参数组合的确定方法同上, 结果如表 4 所示.

表 3 BP 神经网络 (鉴别氦核) 隐藏层节点核验结果

Table 3. BP neural network (identifying helium) hidden layer nodes verification results.

训练结果	隐藏节点个数								
	5	6	7	8	9	10	11	12	13
迭代次数	20000	20000	20000	25000	27000	20000	20000	20000	20000
算法AUC值	0.5503	0.5045	0.5293	0.5593	0.6329	0.6276	0.6177	0.6142	0.6418
Q品质因子	0.82	0.29	0.58	0.86	1.26	1.25	1.22	1.26	1.34

表 4 BP 神经网络鉴别不同成分最佳超参数组合

Table 4. Optimal hyperparameters of BP neural network identifying different components.

超参数	目标成分				
	质子	氦核	碳氮氧	镁铝硅	铁核
隐藏层节点数	13	11	13	13	11
初始学习率	0.01	0.01	0.01	0.01	0.01
迭代次数	20000	25000	20000	20000	20000

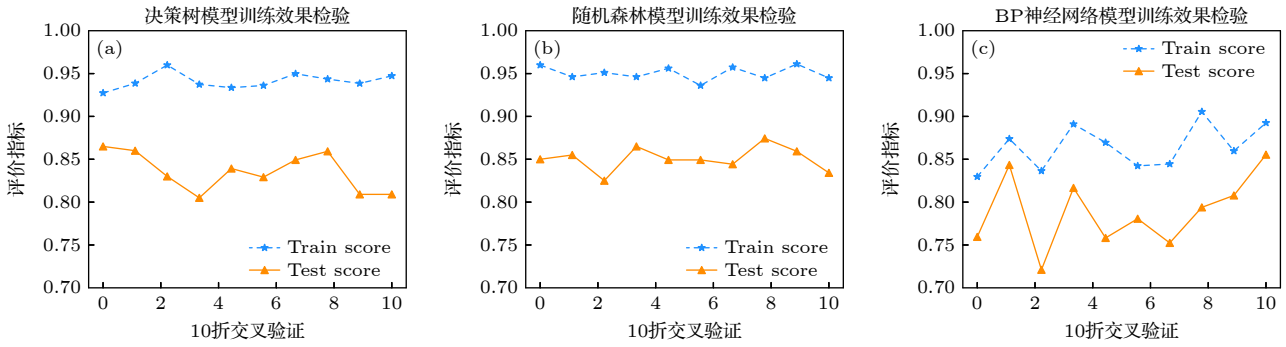

图 4 三种宇宙线鉴别模型鉴别氦十折交叉验证核验图

Fig. 4. Results of three cosmic rays identification models identifying helium using 10-fold cross validation method.

图 4 为 3 种宇宙线粒子鉴别模型鉴别氦核的 10 折交叉验证检验图, 可以看到 10 折交叉验证过程中 3 种模型训练和测试的准确率之差均不超过 0.2, 即 3 种模型均不存在严重的过拟合问题.

3 结果与讨论

本文在训练过程中将目标成分向“0”方向训练, 其他成分向“1”方向训练, 并输出相应的概率. 为了描述 3 种机器学习算法对目标成分 (target) 鉴别的结果, 定义临界值 T_c 来计算目标成分鉴别的纯度 (purity) 和效率 (efficiency), 计算公式如下:

$$\text{Purity} = \frac{N_{\text{target}}\,(T \leqslant T_c)}{N_{\text{all}}\,(T \leqslant T_c)},$$

$$\text{Efficiency} = \frac{N_{\text{target}}\,(T \leqslant T_c)}{N_{\text{target}}\,(\text{All})}. \tag{10}$$

以鉴别目标成分氦核为例, 3 种鉴别模型将粒子种类判定为氦核的概率如图 5 所示, 综合考虑氦核纯度及效率后本文选择临界值 T_c 为 0.5, 即: 1) 在 BP 神经网络鉴别模型中, $T \leqslant 0.5$ 时, 氦核鉴别效率及纯度分别为 36.0%, 52.8%; 2) 在决策树鉴别模型中, $T \leqslant 0.5$ 时, 氦核鉴别效率及纯度分别为 83.3%, 80.1%; 3) 在随机森林鉴别模型中, $T \leqslant 0.5$ 时, 氦核鉴别效率及纯度分别为 79.3%, 95.7%; 由此可以看出, 随机森林算法鉴别氦核纯度较高, 达到 94.5%, 鉴别氦核的效率在 79% 左右.

与模型鉴别氦核过程类似, 其他成分鉴别效率及纯度如表 5 所示. 1) 在利用 BP 神经网络鉴别模型和随机森林鉴别模型鉴别各成分时, 重成分 (铁核) 鉴别的效率及纯度较高, 其中神经网络算法效率和纯度分别为 82.8% 和 87.5%, 随机森林鉴别模型鉴别铁核的效率和纯度分别为 91.1% 和 93.5%;

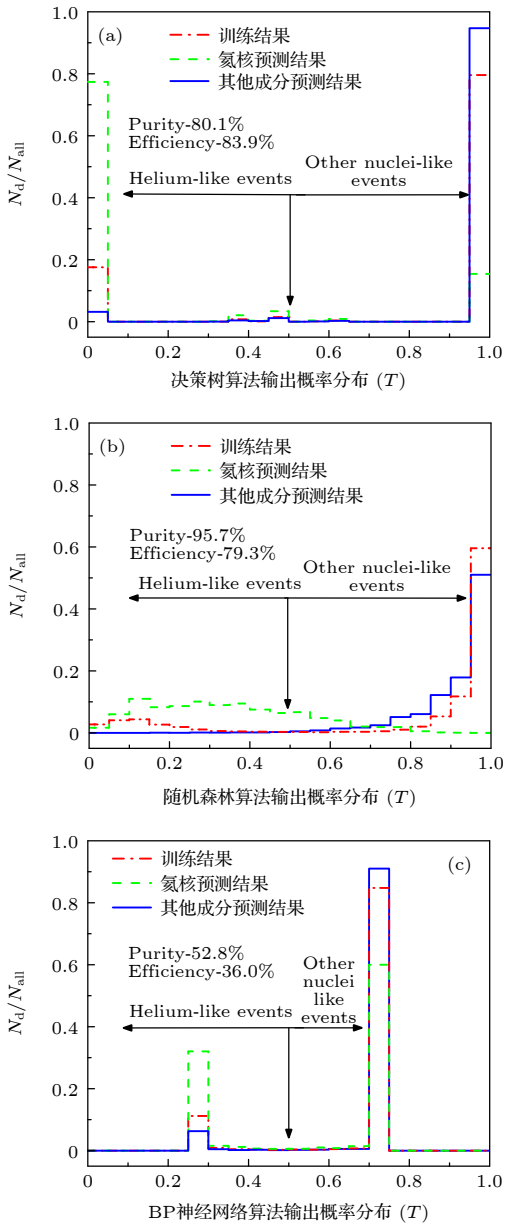

图 5 三种宇宙线粒子鉴别模型鉴别氦核概率分布图

Fig. 5. Probability distribution of three cosmic rays identification models identifying helium.

表 5 三种宇宙线粒子鉴别模型鉴别不同成分效率及纯度

Table 5. Efficiency and purity of three cosmic rays identification models identifying different components.

目标成分	效率/%			纯度/%		
	BP神经网络	决策树	随机森林	BP神经网络	决策树	随机森林
质子	64.9	74.8	75.7	74.4	77.6	91.1
氦核	36.0	83.3	79.3	52.8	80.1	95.7
碳氮氧	10.3	93.4	81.5	64.5	94.8	99.4
镁铝硅	16.9	91.8	78.7	69.9	92.1	95.8
铁核	82.8	88.1	91.1	87.5	88.7	93.5

表 6 三种宇宙线粒子鉴别模型鉴别不同成分 AUC 值及 Q 品质因子

Table 6. AUC and Q quality factor values of three cosmic rays identification models identifying different components.

目标成分	AUC			Q 品质因子		
	BP神经网络	决策树	随机森林	BP神经网络	决策树	随机森林
质子	0.7962	0.8555	0.9247	2.71	3.15	5.42
氦核	0.6418	0.8805	0.9537	1.34	3.75	8.38
碳氮氧	0.5444	0.9612	0.9739	0.87	7.55	20.1
镁铝硅	0.5754	0.9504	0.9531	1.25	6.54	8.39
铁核	0.8751	0.8952	0.9380	2.96	2.97	4.40

2) 在利用决策树鉴别模型鉴别成分时, 对于中成分 (镁铝硅、碳氮氧) 鉴别效率及纯度较高, 效率和纯度均可以达到 90% 以上; 3) 利用 3 种鉴别模型鉴别轻成分 (氦核、质子), 决策树与随机森林鉴别模型鉴别轻成分效率在 74% 以上, 纯度在 77% 以上, 而神经网络鉴别模型鉴别轻成分效率, 尤其是对氦核的鉴别效率与纯度并不高, 对质子鉴别效率与纯度在 64% 以上.

随后, 本文根据各成分鉴别结果得到算法分类效果检验的评价指标 AUC 值与宇宙线研究领域中的品质因子 Q 值 (如表 6 所示), 结果表明: 1) 随机森林算法在各成分判别中纯度均可达到 90% 以上, Q 品质因子较高, 即对宇宙线各成分鉴别能力比其他两种算法要好; 2) 决策树算法在中成分 (镁铝硅、碳氮氧) 鉴别正确率可达 90% 以上, Q 品质因子在 6 以上; 在轻成分和重成分中的鉴别正确率达 85% 以上, Q 品质因子在 3 左右; 3) 神经网络算法在重成分 (铁核) 鉴别中具有一定优势, 判别正确率达到 87%, Q 品质因子为 2.96.

客观来讲, 天顶角、能量以及簇射芯位在阵列中的位置等相关参量也都会受到原初宇宙射线的重建精度的影响, 本文目前在算法建模中采用的参量还比较理想化, 未将以上参量进行综合考量, 下一步我们将在此基础上继续优化和修正机器学习算法模型.

4 结束语

本文将决策树、随机森林、BP 神经网络算法应用在宇宙线粒子分类问题中, 并针对不同算法进行超参数优化调整, 以提高算法判别的正确率及鉴别效率. 实验结果表明, 机器学习算法在宇宙射线粒子成分鉴别领域有较大的应用前景. 目前本文只考虑了 BP 神经网络、决策树和随机森林算法对于宇宙线粒子成分分析的高效率, 还未使用其他算法对宇宙线粒子成分进行分析, 而且训练和模拟所用参数过于理想化, 因此, 下一步研究工作中将加入更接近实验中实际探测的观测量, 进一步优化机器学习算法, 提升粒子鉴别能力, 并将继续深入探索其他机器学习算法在宇宙线粒子鉴别中的应用.

参考文献

[1] Hu H B, Wang X Y, Liu S M 2018 *Chin. Sci. Bull.* **63** 2440 (in Chinese) [胡红波, 王祥玉, 刘四明 2018 科学通报 **63** 2440]

[2] Hu H B, Guo Y Q 2016 *Chin. Sci. Bull.* **61** 1188 (in Chinese) [胡红波, 郭义庆 2016 科学通报 **61** 1188]

[3] Cao Z 2022 *Chin. Sci. Bull.* **67** 1558 (in Chinese) [曹臻 2022 科学通报 **67** 1558]

[4] Cao Z, Chen M J, Chen S Z, Hu H B, Liu C, Liu Y, Ma L L, Ma X H, Shen X D, Wu H R, Xiao G, Yao Z G, Yin L Q, Zha M, Zhang S S 2019 *Acta Astron. Sin.* **60** 3 (in Chinese) [曹臻, 陈明君, 陈松战, 胡红波, 刘成, 刘烨, 马玲玲, 马欣华, 盛祥东, 吴含荣, 肖刚, 姚志国, 尹丽巧, 查敏, 张寿山 2019 天文学报 **60** 3]

[5] Li C, Wang W B, Chen P F 2022 *Sci. China Phys. Mech.* **52** 16 (in Chinese) [李川, 王文博, 陈鹏飞 2022 中国科学: 物理学 力学 天文学 **52** 16]

[6] Zhang F, Liu H, Zhu F R 2022 *Acta Phys. Sin.* **71** 249601 (in Chinese) [张丰, 刘虎, 祝凤荣 2022 物理学报 **71** 249601]

[7] Chen X L 2020 *M. S. Thesis* (Beijing University of Chinese Academy of Sciences) (in Chinese) [陈秀林 2020 硕士学位论文 (北京: 中国科学院大学)]

[8] Yan Y T, Bi W J 2023 *Chin. J. Manag. Sci.* Online First (in Chinese) (in Chinese) [严雨婷, 毕文杰 2023 中国管理科学 网络首发 [2023-02-03]]

[9] Herrera L J, Todero Peixoto C J, Baños O, Carceller J M, Carrillo F, Guillén A 2020 *Entropy* **22** 998

[10] Pang L G, Zhou K, Su N, Petersen H, Stöcker H, Wang X N 2018 *Nat. Commun.* **9** 210

[11] Gao Z P, Wang Y J, Li Q F, Liu L 2022 *Sci. China Phys. Mech.* **52** 252010 (in Chinese) [高泽鹏, 王永佳, 李庆峰, 刘玲 2022 中国科学: 物理学 力学 天文学 **52** 252010]

[12] Luo S J, Han S Z 2023 *J. Chin. Comp. Syst.* Online First (in Chinese) (in Chinese) [骆仕杰, 韩抒真 2023 小型微型计算机系统 网络首发 [2023-03-05]]

[13] Wang Y D, Wang Z H, Zhou R, Chen X L, Qin X, Liu J 2019 *Nucl. Electron. Detect. Technol.* **39** 567 (in Chinese) [王玉东, 王忠海, 周荣, 陈秀莲, 覃雪, 刘军 2019 核电子学与探测技术 **39** 567]

[14] Li W, Long L C, Liu J Y, Yang Y 2022 *Acta Phys. Sin.* **71** 060202 (in Chinese) [黎威, 龙连春, 刘静毅, 杨洋 2022 物理学报 **71** 060202]

[15] Cui J J, Hu Z W, Ren P 2022 *Inf. Sci.* **40** 90 (in Chinese) [崔静静, 胡泽文, 任萍 2022 情报科学 **40** 90]

[16] Li Y N 2018 *Inf. Sci.* **36** 80 (in Chinese) [李勇男 2018 情报科学 **36** 80]

[17] Lin S, Luo W 2019 *Multivar. Behav. Res.* **54** 578

[18] Song S, Park C G 2019 *Sustainability* **11** 6976

[19] Xiao Y, Guo Y H, Li M W, Guo Y S, Sun F 2022 *J. Beijing Normal Univ. (Nat. Sci.)* **58** 261 (in Chinese) [肖燚, 郭亚会, 李明蔚, 付永硕, 孙峰 2022 北京师范大学学报 (自然科学版) **58** 261]

[20] Lu X B, Zhang Y Y, Yang G C, Xing J X 2022 *Inf. Sci.* **41** 1059 (in Chinese) [卢小宾, 张杨燚, 杨冠灿, 行佳鑫 2022 情报学报 **41** 1059]

[21] Pei Y L, Li D D, Xue W X 2020 *Concurr. Comp-Pract E* **32** e5515

[22] Yuan L, Yang X S, Wang B Z 2019 *Acta Phys. Sin.* **68** 170503 (in Chinese) [院琳, 杨雪松, 王秉中 2019 物理学报 **68** 170503]

[23] Zhao Z Y, Liu Y F, Liu S C, Ma H B 2023 *J. Beijing Univ. Aeronaut. Astronaut (Soc. Sci. Ed.)* Online First (in Chinese) [赵振宇, 刘宇帆, 刘善存, 马海波 2023 北京航空航天大学学报 (社会科学版) 网络首发 [2023-03-05]]

Application of machine learning in cosmic ray particle identification*

Liu Ye [1)] Niu He-Ran [1)] Li Bing-Bing [2)] Ma Xin-Hua [3)4)] Cui Shu-Wang [2)†]

1) (*School of Management Science and Engineering*, *Hebei University of Economics and Business*, *Shijiazhuang* 050061, *China*)

2) (*College of Physics*, *Hebei Normal University*, *Shijiazhuang* 050024, *China*)

3) (*Key Laboratory of Particle Astrophysics*, *Institute of High Energy Physics*, *Chinese Academy of Sciences*, *Beijing* 100049, *China*)

4) (*TIANFU Cosmic Ray Research Center*, *Chengdu* 610000, *China*)

(Received 7 March 2023; revised manuscript received 8 April 2023)

Abstract

Machine learning algorithms can learn the rules and patterns of big data through computers, excavate potential information hidden behind the data, and be widely used to solve classification, regression, clustering, and other problems. Firstly, this paper uses CORSIKA software to simulate the process of cosmic ray cascade shower in the atmosphere, generating information such as the initial energy, zenith angle, azimuth angle of cosmic ray particles. Then, this paper uses the Geant4 toolkit to conduct thermal neutron detector response simulation, generating 4000 particles in each of proton, helium, CNO, MgAlSi and iron. Based on the experimental simulation data of thermal neutron detector, this paper constructs machine learning models for identifying cosmic ray particles by using decision tree (DT), random forest (RF) and BP neural network (BP NN) respectively. For each particle, all the machine learning algorithms are used for model training based on the simulation data. The cross grid search method is used to adjust the hyper parameters of each machine learning algorithm. The AUC value and Q quality factor value of each algorithm are used as evaluation indexes for particle composition identification. The AUC value is a general indicator for evaluating algorithm performance in machine learning and the Q quality factor value is an evaluation index commonly used in the field of high energy physics. The Experimental results show that different machine learning models have great influence on particle prediction accuracy, and the random forest cosmic ray particle identification model has sufficient accuracy and generalization capability. In the test, the decision tree algorithm adjusted by cross grid search method is sensitive to the medium components (CNO and MgAlSi). The AUC values of the algorithm are all above 0.95 and the Q quality factor values are all above 6. The random forest algorithm adjusted by the cross grid search method has the best effect on the identification of cosmic ray particles. The AUC values of the algorithm are all more than 0.92 and the Q quality factor values are all more than 4. The BP neural network algorithm is only sensitive to proton and iron. This study provides a new method and selection for identifying and screening the cosmic ray particles and it also provides a new idea for the following measurement of cosmic ray energy spectrum by thermal neutron detector.

Keywords: cosmic rays, particle identification, machine learning, random forest

PACS: 02.70.–c, 95.55.Vj **DOI:** 10.7498/aps.72.20230334

* Project supported by the National Natural Science Foundation of China (Grant Nos. 11905043, U2031103), the Department of Education Project Hebei Province, China (Grant No. KCJSZ2022036), and the Hebei University of Economics and Business Graduate Innovation Funding Project, China (Grant No. XYCX202333).

† Corresponding author. E-mail: cuisw@hebtu.edu.cn

物理学报 Acta Physica Sinica

Chinese Physical Society

Institute of Physics, CAS

机器学习在宇宙线粒子鉴别中的应用

刘烨　牛赫然　李兵兵　马欣华　崔树旺

Application of machine learning in cosmic ray particle identification

Liu Ye　Niu He-Ran　Li Bing-Bing　Ma Xin-Hua　Cui Shu-Wang

引用信息 Citation: Acta Physica Sinica, 72, 140202 (2023)　DOI: 10.7498/aps.72.20230334

在线阅读 View online: https://doi.org/10.7498/aps.72.20230334

当期内容 View table of contents: http://wulixb.iphy.ac.cn
