第二届 GH 数据分析排位赛说明文档

团队: 燕园情 作者: 刘汉青

2020 年 8 月 3 日

1 问题回顾

在本问题涉及的实验装置中，粒子在液体闪烁体中高速运动导致闪烁体发出荧光，在实验装置外围的 30 个光电倍增管（PMT）将光信号转化为电信号进行输出。通过分析光电倍增管的电压波形数据“Waveform”，与由波形计算得到的光电子入射时间“PEGuess”来推断入射粒子的种类是 α 还是 β。

2 算法原理

本团队在比赛中主要使用了多种机器学习算法，包括 Boosting 算法中的 XGBoost 与 Lightgbm，以及深度学习算法中的卷积神经网络（CNN）与长短期记忆神经网络（LSTM）对粒子的特征进行分辨。搭建好 XGBoost、Lightgbm 与 CNN+LSTM 三个分类器之后，将三者的结果进行整合，作为最终结果。

图 1: 算法原理

3 特征工程

3.1 预处理

特征工程的第一步是对原始数据进行预处理。

原始波形数据有较明显的噪声，并且有一个约 973 单位大小的背景。处理过程中，首先，消除背景数据，通过用数据前 10ns 平均值减去原始数据得以实现；之后，平滑去噪，通过将每一个数据用附近 10 个点数据平均进行代替实现。

图 2: 波形预处理

由于 PMT 的物理特性，不同事件的原始光电子到达时间没有统一的基准点，不能直接比较。我们将每个事件接收到 10% 光电子的时间设置为基准时间，对“PEGuess”进行时间对齐。

3.2 特征提取

对于 XGBoost 与 Lightgbm 两个 Boosting 算法分类器，我们进行了特征提取。通过数据分析与运算，我们一共提取了 137 种特征，分为以下四个部分：

1. **光子数随时间的变化**
   α 粒子电荷大，质量大，发光时间常数长，衰减慢；而 β 粒子产生光子密度小，发光常数短，所以可利用光子数随时间变化进行判别。我们提取了 PEGuess 从-20ns 到 240ns 每 5ns 接收到的粒子数，共 51 组数据作为第一部分特征。

2. **切伦科夫光子**
   由于 β 粒子静质量小，初始速度超过介质光速，所以在刚开始的一段时间会发出切伦科夫光子，而 α 粒子无此特性。同时切伦科夫光子具有方向性，分布不均匀，所以可以用来进行分辨。我们提取了每个事件从-10 到 0ns 各个 PMT 上接收光子数，共 30 组数据作为第二部分特征。

3. **消除粒子位置偏差**
   决赛数据 PMT 阈值为 10，光子在各个 PMT 上分布不均。许多 Event 入射粒子初始位置偏离中心很远，导致闪烁体发出的散射光子射出位置也偏离中心很远，具有一定的方向性，对切伦科夫光子的识别造成干扰。我们提取了每个事件各个 PMT 上在整个 Event 过程中接收光子数共 30 组数据作为第三部分特征。

4. **波形的傅里叶频谱**
   与 粒子波形不但在时域有区别，在傅里叶频域也有区别。对波形进行离散快速傅里叶变换（FFT）取实部之后对数据集中所有波形取平均，可以看出两种粒子的傅里叶频谱有很大差别：

图 3: 傅里叶频谱

我们提取了单个 Event 波形离散快速傅里叶变换（FFT）频谱实部前 26 位作为第四部分特征。

将这四部分总共 137 种特征构成的数据集与标签输入两种 Boosting 模型进行学习，就得到了两个 Boosting 分类器。

而卷积神经网络、长短期记忆神经网络等深度学习算法通过模拟人脑学习过程进行分类运算，无需人为提取特征。所以对于 CNN+LSTM 分类器，我们直接将预处理之后的波形数据输入进行学习即可。

4 模型优化

在搭建好基本模型之后我们通过不同角度对模型进行了改进，主要包括以下两个方面。

4.1 调整参数

我们对于 boosting 算法的最大深度（max_depth）、学习速率（learning_rate），深度学习算法的卷积核数（filters），优化器（optimizer）、学习速率（learning_rate）、学习次数（epoch）不断进行手动调整，以期获得更好的效果。

4.2 模型融合

我们最后设计出了 xgboost、lightgbm 与 CNN+LSTM 三种分类器，通过将三个分类器结果进行综合得到最终结果：

1. **加权平均**
   将三种分类器的结果按各自 AUC 值进行加权平均，得到最终结果。

2. **Stacking**
   将三种分类器的预测结果作为三种特征，利用 lightgbm 重新进行学习，得到最终结果。

以下是实验结果：

图 4: 实验结果

5 亮点与不足

亮点 1：使用机器学习算法

本团队使用 XGBoost、Lightgbm 等基于决策树的机器学习算法与神经网络深度学习算法，通过计算机强大的运算能力，自动提取特征与粒子种类的关系，解决了大量原本需要人工推导的部分，大大方便了科研工作。

亮点 2：算法简洁

本团队使用了 python 语言的 Numpy、Pandas 等数据分析第三方库与 SkLearn、TensorFlow 等机器学习第三方库，通过大量封装函数，大大精简代码，实际训练过程代码不超过五十行，一方面减少了科研人员的工作，一方面降低了数据分析学习门槛。

不足 1：算法结构优化不多

由于本团队的物理专业背景，我们在实验过程中更加关注特征提取等偏物理方面的工作。而对算法优化改进不多，如实验过程中很少调整模型参数。

不足 2：切伦科夫光利用较少

由于设备限制，本团队的模型对切伦科夫光空间特性利用较少。

$$\bigcap_{k=1}^{\infty}$$