机器学习如何帮助天文学家监视望远镜状态

我们头顶的天空已经被观测了几千年。最初人们肉眼观察星星，望远镜出现后，人类的视野越来越广阔。为了得到更清晰的观测，人们甚至把望远镜送入了太空，可以看到 100 万光年以外的星星。

为了探索更深远的宇宙，现代望远镜越来越庞大复杂。以射电望远镜为例，为了提高空间解析度和信号敏感性，望远镜的天线越来越大，甚至组成庞大的天线群来协同完成观测。例如，位于智利查南托高原的射电望远镜就包括 66 座天线，最大的天线直径达 12 米[1]。

这些大型观测设备每天都在瞭望星空，每时每刻都在产生海量数据。这些数据中固然包含丰富的信息，但已经不是人用肉眼可以分析和理解的了。或者说，当下的天文学研究已经进入了大数据时代，必须有相对应的研究工具才能从这些海量数据中发现有价值的线索，而这正是机器学习所擅长的。归因于此，近年来机器学习在天文学研究中异军突起[2]，特别是深度学习方法，因其强大的数据学习能力受到青睐，广泛应用在光谱分析，新星检测，星系归类等任务中[3,4]。

一个有趣的例子是荷兰科学家发表在 2020 年 3 月英国皇家天文学会月刊上的一篇文章，用神经网络来监控射电望远镜的工作状态[4]。为什么要做这个工作呢？这是因为每天采集到的数据实在是太多了，多到连望远镜工作异常都不容易发现。这就带来一个非常严重的问题，如果连仪器是否正常工作都不知道，如何保证得到的数据是可信的，又如何依赖这些数据得到可靠的结果呢？

为了解决这个问题，研究者采用了一种称为变分自编码器（Variational Auto Encoder, VAE）的神经网络，将望远镜观察到的数据投影到一个二维空间，如果设备发生异常，数据将在这个二维空间中产生偏移，这样就能及早发现问题。图 1 是该 VAE 的结构图，其基本思路是从观察到的数据中提取出幅度谱（Input Magnitude）和相位谱（Input Phase），将他们同时送入编码器，通过一系列变换，得到一个二维空间中的嵌入向量（Embedding），再经过一个解码器还原出原始幅度谱和相位谱。模型的训练准则是使输出端尽可能还原输入端的所有信息，包括幅度谱和相位谱。由于嵌入向量只有二维，这一训练将迫使嵌入向量尽可能保留输入数据中的重要信息。正因为如此，可以通过这一嵌入向量来观察数据中可能出现的异常。

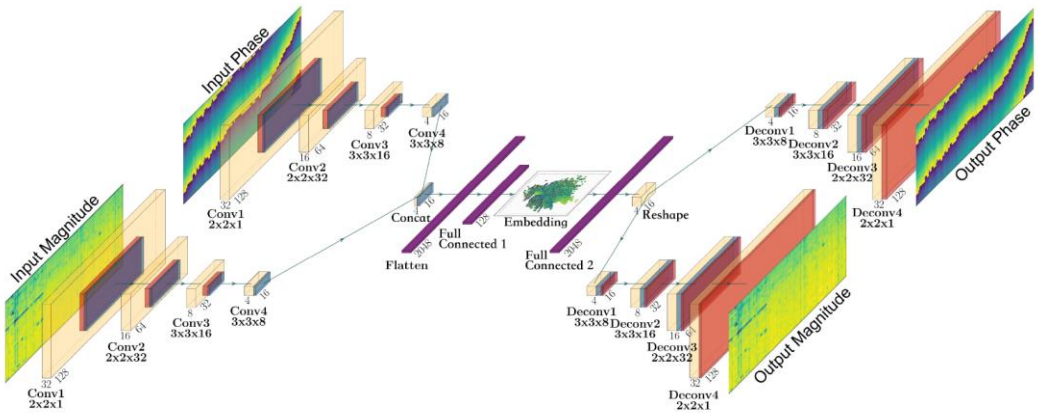

图 1：利用 VAE 将观察到的数据映射到二维空间中的嵌入向量。

图 2 给出了一个模拟实验结果，其中每个点是一个观察数据，每种颜色代表一种可能的异常组合，如射电频域干扰或高斯噪声等。可以看到，不同异常状态可以清晰地反映在嵌入向量组成的二维空间中。反过来，通过观察这一空间即可定位可能出现的数据异常。

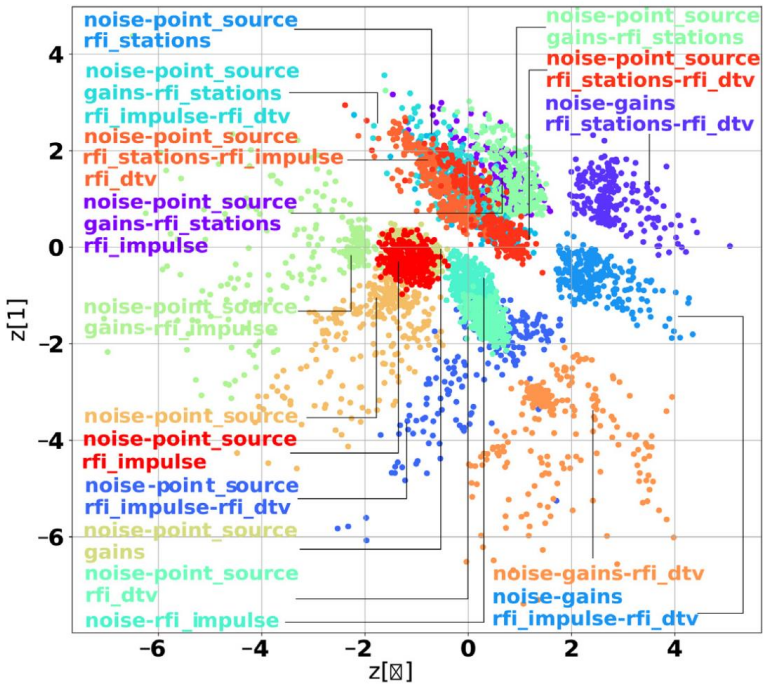

图 2：不同特性的数据在嵌入向量空间的分布。

图 3 是一个称为 LOFAR 的实际数据集上的嵌入向量分布，其中左图展示的是每个嵌入位置对应的幅度谱，而右图展示的是每个嵌入位置对应的相位谱。可以看到，不同模式的数据被映射到了二维空间中的不同位置，而同一位置的数据模式具有相似性。这一结果表明该方法确实可以为天文学家提供一种直观的工具，不仅可以监视设备运行的状态，还可能有更多有

价值的应用。

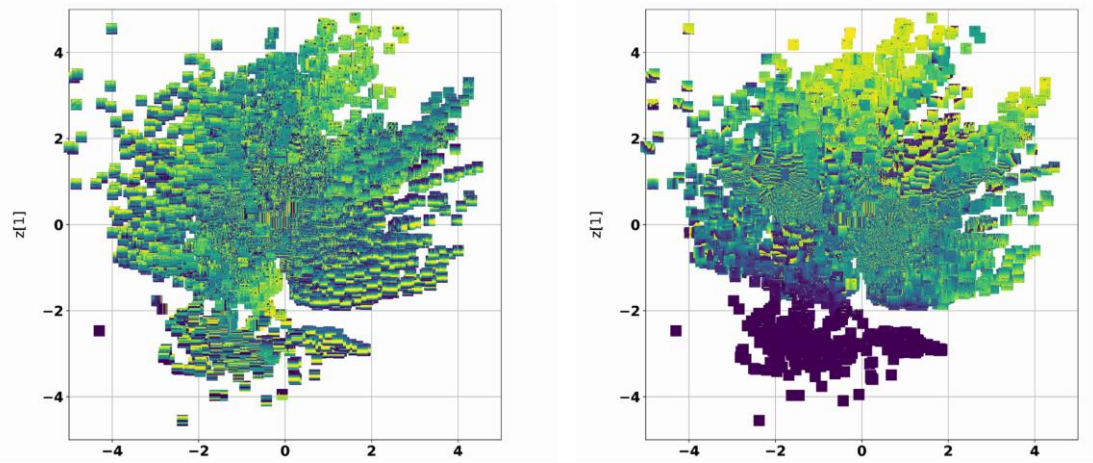

图 3：LOFAR 数据的二维嵌入向量。左图中每个点是位于该位置数据的幅度谱，右图中每个点是位于该位置数据的相位谱。

[1]最强大射电望远镜亮相由 66 座天线构成，https://epaper.qlwb.com.cn/qlwb/content/20130314/ArticelA30002FM.htm
[2] Baron D. Machine learning in astronomy: A practical overview[J]. arXiv preprint arXiv:1904.07248, 2019.
[3] Henry W. Leung1 and Jo Bovy, Deep learning of multi-element abundances from high-resolution spectroscopic data, MNRAS, 2018.
[4] Mesarcik et al, Deep learning assisted data inspection for radio astronomy, MNRAS, 2020.