中国科学院大学
University of Chinese Academy of Sciences

使用集成强化方式处理不平衡数据

曾鴻

中国科学院国家天文台

2023 年 10 月 8 日, 北京怀柔

中国科学院大学
University of Chinese Academy of Sciences

目录

中国科学院大学
University of Chinese Academy of Sciences

分类周期变星存在的问题

不同类别周期变星的样本量存在高度不平衡。

中国科学院大学
University of Chinese Academy of Sciences

现有的解决方法

- 传统机器学习算法:SMOTE、Self-paced Ensemble 等
- 数据生成: 添加噪声、高斯过程、深度生成模型

中国科学院大学
University of Chinese Academy of Sciences

数据层面

利用高斯过程生成小样本量类别的合成光变曲线。

中国科学院大学
University of Chinese Academy of Sciences

模型层面

- RNN-based 神经网络
- RNN + CNN 混合结构神经网络

中国科学院大学
University of Chinese Academy of Sciences

训练优化

采用 bagging-like 方式组织数据增强和模型训练，减轻过拟合。

中国科学院大学
University of Chinese Academy of Sciences

取得的主要结果

- 在 CRTS 数据集上 Macro F1 得分达到 0.75。
- 总体精度达到 86.2%。

中国科学院大学
University of Chinese Academy of Sciences

与现有方法的不同

- 使用高斯过程而不是 SMOTE 生成合成数据。
- 尝试了不同的神经网络架构。
- 采用了 bagging-like 的模型集成方法。

中国科学院大学
University of Chinese Academy of Sciences

未来展望

- 继续优化小样本类别分类。
- 探索更高效的模型和训练方式。
- 在其他数据上验证所提出的方法。

中国科学院大学
University of Chinese Academy of Sciences

参考文献 I

[1] KANG Z, ZHANG Y, ZHANG J, et al. Periodic Variable Star Classification with Deep Learning: Handling Data Imbalance in an Ensemble Augmentation Way[J]. Publications of the Astronomical Society of the Pacific, 2023, 135(1051): 094501.

谢谢