面向多类不平衡学习的一对多海林格距离决策树研究

董明刚[1,2]，刘明[1,2]，敬超[1,2,3]

[1]桂林理工大学信息科学与工程学院，中国桂林市，541004

[2]广西嵌入式技术与智能系统重点实验室，中国桂林市，541004

[3]桂林电子科技大学广西可信软件重点实验室，中国桂林市，541004

摘要：由于传统机器学习方法对偏斜分布很敏感，且未考虑多类不平衡问题的特点，多类偏斜分布对机器学习算法来说是一个巨大挑战。为解决这一问题，提出一种新的基于一对多的海林格距离（OAHD）决策树分割准则。OAHD主要由两部分组成。首先，将一对多思想集成到OAHD的海林格距离计算过程中，从而对海林格距离决策树进行扩展，使其能解决多类不平衡问题。其次，针对多类不平衡问题，考虑了不同类的分布和数量，设计了改进的基尼系数。此外，对OAHD的性质进行理论证明，包括偏斜不敏感性和在决策树中寻找更纯节点的能力。最后，从基于进化学习的知识抽取（KEEL）和加州大学欧文分校（UCI）数据库中收集20个公开的真实不平衡数据集进行实验。实验结果表明，与其他5种常用决策树相比，OAHD在精度、F值，和多类别接收者操作特征曲线下面积（MAUC）上有显著优势。此外，使用了Friedman和Nemenyi检验，统计结果表明OAHD优于其他5种决策树。

关键词：决策树；多类不平衡学习；节点划分准则；海林格距离；一对多技术

https://doi.org/10.1631/FITEE.2000417