计算机应用 2006 26 (08): 1894-1897 ISSN: 1001-9081 CN: 51-1307/TP

数据库与软件技术

不均衡数据集学习中基于初分类的过抽样算法

韩慧[1]; 王路[2]; 温明[2]; 王文渊[2];[2]

清华大学自动化系[1]

收稿日期 2006-3-1 修回日期 网络版发布日期 2006-8-1 接受日期

摘要 为了有效地提高不均衡数据集中少数类的分类性能，提出了基于初分类的过抽样算法。首先，对测试集进行初分类，以尽可能多地保留多数类的有用信息；其次，对于被初分类预测为少数类的样本进行再次分类，以有效地提高少数类的分类性能。使用美国加州大学欧文分校的数据集将基于初分类的过抽样算法与合成少数类过抽样算法、欠抽样方法进行了实验比较。结果表明，基于初分类的过抽样算法的少数类与多数类的分类性能都优于其他两种算法。

关键词 不均衡数据集 过抽样 欠抽样

分类号

DOI:

对应的英文版文章：6020959

通讯作者：
韩慧 hanh01@mails.tsinghua.edu.cn
作者个人主页：韩慧 王路 温明 王文渊