一种用于不平衡学习分类的新型交叠最小化 SMOTE 算法

何玉林[1,2]，路璇[2]，Philippe FOURNIER-VIGER[2]，黄哲学[1,2]
[1] 人工智能与数字经济广东省实验室（深圳），中国深圳市，518107
[2]深圳大学计算机与软件学院，中国深圳市，518060

摘要：合成少数类过采样技术（SMOTE）是不平衡学习领域的经典算法之一，用于减轻类别不平衡对构建分类器的影响。在过去20年中，有上百个基于SMOTE的变体算法被提出。SMOTE及其变体算法通过在原始样本空间中对少数类样本进行插补来平衡数据集，以减轻类别不平衡的不利影响。这种方法在许多情况下表现良好，但当合成样本落入类别之间的交叠区域时，分类器训练的复杂性会增加，进而影响分类器的泛化能力。为解决这一问题，本文提出一种基于交叠最小化的少数类样本生成算法（Overlapping Minimization SMOTE，OM-SMOTE），用于解决二元不平衡分类问题。OM-SMOTE首先通过平衡样本编码和分类器泛化之间的权衡，将原始样本点映射到更加线性可分的样本空间。然后，OM-SMOTE采用一系列复杂的少数类样本点插补规则，使合成样本尽可能远离类别交叠的区域。本文基于32个真实不平衡数据集进行了大量实验，验证了OM-SMOTE算法的有效性。实验结果表明，相对于其他11种先进的基于SMOTE的过采样算法，OM-SMOTE生成的少数类样本点能显著提高朴素贝叶斯、支持向量机、决策树和逻辑回归等分类器的性能。这证明了OM-SMOTE支持训练高质量不平衡分类器的可行性。OM-SMOTE的实现在GitHub平台上（https://github.com/luxuan123123/OM-SMOTE/）公开共享。

关键词：不平衡分类；合成少数类过采样技术；多数类样本；少数类样本；泛化能力；交叠最小化
https://doi.org/10.1631/FITEE.2300278