山东大学学报(工学版) 2011, 41(3) 7-11 DOI: ISSN: 0412-1961 CN: 21-1139

机器学习与数据挖掘

面向不平衡问题的集成特征选择

李霞1,王连喜2,蒋盛益1

1. 广东外语外贸大学信息学院, 广东 广州 510006;
2. 广东科贸职业学院商贸系, 广东 广州 510640

摘要:

传统的特征选择方法基本上是以精度为优化目标,没有充分考虑数据样本类别分布倾斜性,在数据分布不平衡的数据集上性能表现不理想。在不平衡数据集上通过有放回的抽样方法独立地从数据集大类样本集中随机抽取多个样本子集,使每次随机抽取的样本数量与小类样本数量一致,然后将各抽取的样本子集分别与小类样本集组合成多个新的训练样本集。对多个新样本集的特征子集以集成学习的方式采用投票机制进行投票,数据集的最终特征子集以得票数目超过半数的特征共同组合而成。在UCI不平衡数据集上的实验结果显示,提出的方法表现出了较好的性能,是一种能够处理不平衡问题的有效特征选择方法。

关键词: 不平衡数据集 特征选择 集成学习 抽样

Ensemble learning based feature selection for imbalanced problems

LI Xia1, WANG Lian-xi2, JIANG Sheng-yi1

1. School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006, China;
2. Department of Business and Trade, Guangdong Vocational College of Science and Trade, Guangzhou 510640, China

Abstract:

The traditional feature selection methods are basically aimed for getting the optimal accuracy without full consideration of the data distribution, which can not achieve promising results on imbalanced datasets. A new feature selection method was proposed based on the data distribution modification for imbalanced data sets. This approach could modify data distribution many times by sampling with replacement. The instances of large classes were equal to the minor class samples in each new dataset. Finally, the final selected features were generated by voting mechanism for ensemble learning, which could combine the selected features by receiving more votes than half from all the new training datasets. Experimental results on several UCI datasets showed that the proposed method was an effective feature selection approach for imbalance problems.

Keywords: imbalanced data feature selection ensemble learning sampling

收稿日期 2011-02-01 修回日期 网络版发布日期

DOI:

基金项目:

国家自然科学基金资助项目(61070061);广东省自然科学基金资助项目(9151026005000002);广东省高层次人才资助项目

通讯作者:

作者简介: 李霞(1976-),女,江西乐平人,讲师,硕士,主要研究方向为数据挖掘.E mail: shelly-lx@126.com

作者Email:

PDF Preview

参考文献:

本刊中的类似文章

1．孙晓燕1,2,张化祥1,2*,计华1,2.基于AdaBoost的欠抽样集成学习算法[J]. 山东大学学报(工学版), 2011,41(4): 91-94