

软件技术与数据库

一种改进的少数类样本过抽样算法

许丹丹¹, 蔡立军¹, 王勇²

(1. 西北工业大学理学院, 西安 710129; 2. 西北工业大学计算机学院, 西安 710072)

摘要: 针对偏斜数据集的分类问题, 提出一种改进的少数类样本过抽样算法(B-ISMOTE)。在边界少数类实例及其最近邻实例构成的 n 维球体空间内进行随机插值, 以此产生虚拟少数类实例, 减小数据的不均衡程度。在实际数据集上进行实验, 结果证明, 与 SMOTE 算法和 B-SMOTE 算法相比, B-ISMOTE 算法具有较优的分类性能。

关键词: 偏斜数据集 分类 过抽样 虚拟实例 n 维球体空间

Improved Over-sampling Algorithm of Minority Class Sample

XU Dan-dan¹, CAI Li-jun¹, WANG Yong²

(1. School of Science, Northwestern Polytechnical University, Xi'an 710129, China; 2. School of Computer, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Aiming at the classification of the skewed dataset, this paper proposes an improved over-sampling algorithm of minority class sample, named B-ISMOTE. It improves the data unbalanced distribution of degree through randomized interpolation to produce virtual minority class instances in the sphere space, which constitute of the borderline minority class instances and its nearest neighbor. Experimental results on the real datasets show that compared with SMOTE algorithm and B-SMOTE algorithm, B-ISMOTE algorithm has better classification performance.

Keywords: skewed dataset classification over-sampling virtual instance n dimension sphere space

收稿日期 2011-07-18 修回日期 网络版发布日期 2012-02-20

DOI: 10.3969/j.issn.1000-3428.2012.04.022

基金项目:


国家自然科学基金资助项目(60873196)

通讯作者:

作者简介: 许丹丹(1984—), 女, 硕士研究生, 主研方向: 偏斜数据挖掘; 蔡立军、王勇, 副教授

通讯作者 E-mail: xudandan@mail.nwpu.edu.cn

参考文献:

[3] Gao Jing.[J].Fan Wei, Han Jiawei, et al. A General Framework for Mining Concept-drifting Data Streams with Skewed Distri- butions[C]//Proc. of SDM'07. Minneapolis, USA: [s. n.].2007, : - 

[6] Han Hui.[J].Wang Wenyuan, Mao Binghuan. Borderline-SMOTE: A New Over-sampling

扩展功能

本文信息

- ▶ Supporting info
- ▶ PDF(264KB)
- ▶ [HTML] 下载
- ▶ 参考文献[PDF]
- ▶ 参考文献

服务与反馈

- ▶ 把本文推荐给朋友
- ▶ 加入我的书架
- ▶ 加入引用管理器
- ▶ 引用本文
- ▶ Email Alert
- ▶ 文章反馈
- ▶ 浏览反馈信息

本文关键词相关文章



- ▶ 偏斜数据集
- ▶ 分类
- ▶ 过抽样
- ▶ 虚拟实例
- ▶ n 维球体空间

本文作者相关文章

- ▶ 许丹丹
- ▶ 蔡立军
- ▶ 王勇

PubMed

- ▶ Article by Hu, D. D.
- ▶ Article by Ca, L. J.
- ▶ Article by Wang, Y.

- [7] 王和勇, 樊泓坤, 姚正安. SMOTE和Biased-SVM 相结合的不平衡数据分类方法[J]. 计算机科学. 2008, 35 (5): 174-176 
- [8] 韩 慧, 王 路, 温 明, 等. 不均衡数据集学习中基于初分类的过抽样算法[J]. 计算机应用. 2006, 26 (8): 1894-1897 
- [9] 杜 娟, 衣治安, 周 颖. 基于聚类 and 遗传交叉的少数类样本生成方法[J]. 计算机工程. 2009, 35(22): 182-184 [浏览](#)

本刊中的类似文章

- 1. 钱琳, 秦亮曦. 按需系综的数据流分类算法研究 [J]. 计算机工程, 2012, 38(5): 62-63, 69
- 2. 刘建伟, 李双成, 罗雄麟. 迭代再权q范数正则化LS SVM分类算法[J]. 计算机工程, 2012, 38(3): 166-168
- 3. 华秀秀, 马莉. 黑色素瘤表面不均匀性的轮廓描述与分析[J]. 计算机工程, 2012, 38(3): 196-199
- 4. 冯筠, 李刚, 孙霞, 冯宏伟. 一种面向教学的知识点库自动生成方法[J]. 计算机工程, 2012, 38(2): 201-203
- 5. 张伟松, 高智英. 快速多分类器集成算法研究[J]. 计算机工程, 2012, 38(2): 178-180
- 6. 汪海滨, 查代奉, 龙俊波. α 稳定分布噪声下的空间时频DOA估计[J]. 计算机工程, 2012, 38(2): 284-284
- 7. 张玉培, 孔敏, 翟素兰, 罗斌. 基于镜头标记与动态滑动窗口的视频摘要生成[J]. 计算机工程, 2012, 38(2): 256-258
- 8. 李道申, 刘勇. 基于本体的Deep Web数据源发现方法[J]. 计算机工程, 2012, 38(04): 52-54
- 9. 唐英, 李应珍. 线性支持向量机多类分类器几何构造方法[J]. 计算机工程, 2012, 38(04): 152-154
- 10. 李鲲鹏, 兰巨龙. 基于TCAM的高效浮动关键词匹配算法[J]. 计算机工程, 2012, 38(04): 269-271

文章评论

反馈人	<input type="text"/>	邮箱地址	<input type="text"/>
反馈标题	<input type="text"/>	验证码	<input type="text" value="9663"/>
<input type="text"/>			