系统工程理论与实践
Systems Engineering - Theory & Practice
邮发代号: 2-305
ISSN 1000-6788
CN 11-2267/N

系统工程理论实践 » 2012, Vol. » Issue (8): 1784-1792 DOI:

LSVDD: 基于局部支持向量数据描述的稀有类分析算法

熊海涛[1], 吴俊杰[1], 刘鲁[1], 李明[2]

1. 北京航空航天大学 经济管理学院, 北京 100191;
2. 中国石油大学 工商管理学院, 北京 102249

LSVDD: Rare class analysis based on local support vector data description

XIONG Hai-tao[1], WU Jun-jie[1], LIU Lu[1], LI Ming[2]

1. School of Economics and Management, Beihang University, Beijing 100191, China;
2. School of Business Administration, China University of Petroleum, Beijing 102249, China

- 摘要
- 参考文献
- 相关文章

全文: PDF (856 KB) HTML (1 KB) 输出: BibTeX | EndNote (RIS) 背景资料

摘要 在单类支持向量数据描述算法的基础上, 提出了一种基于局部支持向量数据描述的稀有类分析算法: LSVDD, 能够处理存在类重叠的类不平衡问题. 该算法利用支持向量数据描述算法对各类样本分别进行单类学习, 从而获得单类模型; 然后对单类模型的概念重叠区域使用属性选择进一步进行局部单类学习, 最后得到综合分类模型. 在仿真数据集和UCI数据集上的实验结果表明, LSVDD能够有效和稳定地提高稀有类分析精度.

关键词: 数据挖掘 稀有类分析 支持向量数据描述 属性选择

Abstract: As a hot topic in data mining society, rare class analysis (RCA) has been widely used in various application domains including financial fraud detection, network intrusion detection, facility failure diagnosis, etc. However, it is not until recently that researchers have realized the impact of complex data structures to the RCA problem. We propose a local support vector data description algorithm LSVDD for RCA based on SVDD, which has the ability to handle class imbalance problem with the presence of class overlaps. Specifically, LSVDD firstly uses SVDD to get one-class classification model for each class and finds the concept overlapping regions between different classes. Then, the regions are locally trained using SVDD again after attribute selections. Finally, the models for non-overlapping and overlapping regions are combined to form a complete RCA model. Experimental results on artificial and real-world UCI data sets demonstrate that LSVDD can improve the performances of RCA stably and effectively.

Key words: data mining rare class analysis support vector data description attribute selection

收稿日期: 2010-06-04;

基金资助:国家自然科学基金(70901002, 90924020);高等学校博士学科点专项科研基金(200800060005, 20091102120014)

服务
- 把本文推荐给朋友
- 加入我的书架
- 加入引用管理器
- E-mail Alert
- RSS

作者相关文章
- 熊海涛
- 吴俊杰
- 刘鲁
- 李明

引用本文:

熊海涛,吴俊杰,刘鲁等. LSVDD: 基于局部支持向量数据描述的稀有类分析算法[J]. 系统工程理论实践, 2012, (8): 1784-1792.

XIONG Hai-tao,WU Jun-jie,LIU Lu et al. LSVDD: Rare class analysis based on local support vector data description[J]. Systems Engineering - Theory & Practice, 2012, (8): 1784-1792.

[1] Weiss G M. Mining with rarity: A unifying framework[J]. SIGKDD Explorations, 2004, 6(1): 7-19.

[2] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining[M]. New York: Addison Wesley, 2005: 95-98.

[3] Wu J, Xiong H, Chen J. COG: Local decomposition for rare class analysis[J]. Data Mining and Knowledge Discovery, 2010, 20(2): 191-220.

[4] He H, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[5] 刘叶青, 刘三阳, 谷明涛. 多项式光滑的半监督支持向量分类机[J]. 系统工程理论与实践, 2009, 29(7): 113-118.Liu Y Q, Liu S Y, Gu M T. Improved learning algorithm with transductive support vector machines[J]. Systems Engineering -- Theory & Practice, 2009, 29(7): 113-118.

[6] Cortes C, Vapnik V N. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[7] Tax D, Duin R. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.

[8] Tax D, Duin R. Growing a multi-class classifier with a reject option[J]. Pattern Recognition Letters, 2008, 29(10): 1565-1570.

[9] 徐晶, 石端银, 张亚江, 等. 基于聚类和SVDD 的单类入侵检测模型[J]. 控制与决策, 2010, 25(3): 441-444.Xu J, Shi D Y, Zhang Y J, et al. Model of IDS based on SVDD and cluster algorithm[J]. Control and Decision, 2010, 25(3): 441-444.

[10] Prati R C, Batista G. Class imbalances versus class overlapping: An analysis of a learning system behavior[C] // Proceedings of the Mexican International Conference on Artificial Intelligence, Berlin: Springer, 2004: 312-321.

[11] Weiss G M, Provost F. Learning when training data are costly: The effect of class distribution on tree induction[J]. Journal of Artificial Intelligence Research, 2003, 19: 315-354.

[12] Visa S, Ralescu A. The effect of imbalanced data class distribution on fuzzy classifiers-experimental study[C]// Proceedings of the IEEE Conference on Fuzzy Systems, Washington: IEEE Press, 2005: 749-754.

[13] Wu J, Xiong H, Wu P, et al. Local decomposition for rare class analysis[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM Press, 2007: 814-823.

[14] Gangemi A, Pisanelli D M, Steve G. An overview of the ONIONS project: Applying ontologies to the integration of medical terminologies[J]. Data and Knowledge Engineering, 1999, 31(2): 183-220.

[15] Garc\'ia V, Mollineda R A, S\'anchez J S. On the k-NN performance in a challenging scenario of imbalance and overlapping[J]. Pattern Analysis & Applications, 2008, 11(3/4): 269-280.

[16] Liu C L. Partial discriminative training for classification of overlapping classes in document analysis[J]. International Journal on Document Analysis and Recognition, 2008, 11(2): 53-65.

[17] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. The Journal of Machine Learning Research, 2003(3): 1157-1182.

[18] Tax D. DD_tools[EB/OL]. http://www-ict.ewi.tudelft.nl/davidt/dd_tools.html.

[19] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30(6): 1145-1159.

[20] Juszczak P, Tax D, Pe E, et al. Minimum spanning tree based one-class classifier[J]. Neurocomputing, 2009, 72(7/9): 1859-1869.

[21] Blake C L, Merz C J. UCI repository of machine learning databases[EB/OL]. http://kdd.ics.uci.edu.

[1] 李永立, 吴冲, 王崑声. 基于图论和信息最大化保留的在线推荐方法[J]. 系统工程理论实践, 2011, 31(9): 1718-1725.

[2] 张永杰;张维;金曦;熊熊. 互联网知道的更多么?------网络开源信息对资产定价的影响[J]. 系统工程理论实践, 2011, 31(4): 577-586.

[3] 郭崇慧, 苏木亚. 基于独立成分分析的时间序列谱聚类方法[J]. 系统工程理论实践, 2011, 31(10): 1921-1931.

[4] 琚春华;郭飞鹏. 基于支持向量机的分布数据挖掘模型DSVM[J]. 系统工程理论实践, 2010, 30(10): 1855-1863.

[5] 李杰;徐勇;王云峰;朱昭贤. 面向个性化推荐的强关联规则挖掘[J]. 系统工程理论实践, 2009, 29(8): 144-152.

[6] 王琛. 基于模糊前沿面的分类方法[J]. 系统工程理论实践, 2009, 29(2): 121-126.

[7] 周复之. 固定收益决策支持系统机理建模与数据挖掘的协同研究[J]. 系统工程理论实践, 2009, 29(12): 38-45.

[8] 崔婧;赵秀娟;宋吟秋. 中日股价序列相似性的比较分析[J]. 系统工程理论实践, 2009, 29(12): 125-133.

[9] 兰秋军;马超群;文风华. 基于共同机制的时间序列关联模式挖掘系统及其应用[J]. 系统工程理论实践, 2004, 24(8): 73-79.

[10] 王自强;冯博琴. 频繁项集的简洁表示方法研究[J]. 系统工程理论实践, 2004, 24(7): 74-81.

[11] 姚敏;沈斌;李明芳. 基于多准则神经网络与分类回归树的电信行业异动客户识别系统[J]. 系统工程理论实践, 2004, 24(5): 78-83.

[12] 张曙红;孙建勋;诸克军. 基于遗传优化的采样模糊C均值聚类算法[J]. 系统工程理论实践, 2004, 24(5): 121-125.

[13] 张喆;常桂然;黄小原. 一种基于遗传算法的多重决策树组合分类方法[J]. 系统工程理论实践, 2004, 24(4): 63-69.

[14] 姚敏;沈斌;易文晟. 基于知识生命期的数据挖掘方法研究[J]. 系统工程理论实践, 2004, 24(2): 74-78.

[15] 李仁璞;王正欧. 规则不确定性的几种度量及其相互关系[J]. 系统工程理论实践, 2004, 24(1): 83-87.