计算机应用 2009, 29(10) 2736-2740 DOI: ISSN: 1001-9081 CN: 51-1307/TP

人工智能

一种新的支持向量机大规模训练样本集缩减策略

朱方[1],顾军华[2],杨欣伟[1],杨瑞霞[3]

1. 河北工业大学
2. 河北工业大学 计算机科学与软件学院
3. 河北工业大学 信息工程学院

摘要： 支持向量机(SVM)在许多实际应用中由于训练样本集规模较大且具有类内混杂孤立点数据，引发了学习速度慢、存储需求量大、泛化能力降低等问题，成为直接使用该技术的瓶颈。针对这些问题，通过在点集理论的基础上分析训练样本集的结构，提出了一种新的支持向量机大规模训练样本集缩减策略。该策略运用模糊聚类方法快速的提取出潜在支持向量并去除类内非边界孤立点，在减小训练样本集规模的同时，能够有效地避免孤立点数据所造成的过学习现象，提高了SVM的泛化性能，在保证不降低分类精度的前提下提高训练速度。

关键词： 支持向量机 点集 模糊C-均值 潜在支持向量 孤立点

New reduction strategy of large-scale training sample set for SVM

Abstract: It has become a bottleneck to use Support Vector Machine (SVM) due to such problems as slow learning speed, large buffer memory requirement, low generalization performance and so on, which are caused by large-scale training sample set and outlier data immixed in the other class. Concerning these problems, this paper proposed a new reduction strategy for large-scale training sample set according to the analysis on the structure of the training sample set based on the point set theory. This new strategy gets the potential support vectors and removes the non-boundary outlier data immixed in the other class by using fuzzy clustering. That can greatly reduce the scale of the training sample set and improve the generalization performance by effectively avoiding over-learning caused by outlier data, and finally speed up learning rate without reducing the classification accuracy.

Keywords: Support Vector Machine (SVM) point set Fuzzy C-Means (FCM) potential support vector outlier

收稿日期 2009-04-16 修回日期 2009-06-01 网络版发布日期 2009-10-28

DOI:

基金项目:

天津市自然科学基金

通讯作者: 杨欣伟

作者简介:

作者Email: yangxinwei@126.com

参考文献：

本刊中的类似文章

1．吴德会;Dehui Wu .一种基于LS-SVM的特征提取新方法及其在智能质量控制中的应用[J]. 计算机应用, 2006,26(10): 2446-2449

2．崔霞 童学锋 黄聪.基于马尔可夫模型和支持向量机的JPEG图像隐写分析[J]. 计算机应用, 2007,27(9): 2140-

2142
3．赵明渊 周明天 许雄基 张渡.基于支持向量机的脑-机接口模式分类和模型参数研究[J]. 计算机应用, 2007,27(2): 337-339
4． 张冰 孔锐 .一种支持向量机的组合核函数[J]. 计算机应用, 2007,27(1): 44-46
5．吴少雄 黄恩洲 .基于支持向量机的控制图模式识别[J]. 计算机应用, 2007,27(1): 61-64
6．唐玉华 杨晓元 张敏情 韩鹏 .多超球面OC-SVM算法在隐秘图像检测中的应用[J]. 计算机应用, 2006,26(12): 2887-2889
7．祁云平 张其善 佟雨兵.基于PSNR与SSIM联合的图像质量评价模型[J]. 计算机应用, 0,(): 503-506
8．江力 胡永祥 .非均匀采样曲线的支持向量机重建[J]. 计算机应用, 2006,26(12): 2832-2834
9．薛欣 贺国平 .基于SVM决策树判别测试点类别的新方法[J]. 计算机应用, 2007,27(1): 84-85
10．王晓丹 郑春颖;吴崇明 .一种新的SVM对等增量学习算法[J]. 计算机应用, 2006,26(10): 2440-2443
11．山艳 须文波 孙俊 .QPSO算法在训练支持向量机中的应用[J]. 计算机应用, 2006,26(11): 2645-2647
12．张长 邱保志 .LDC-mine——基于局部偏差系数的孤立点挖掘算法[J]. 计算机应用, 2007,27(1): 95-97
13．艾武 李红 鲁胜强.基于模糊支持向量机的色素皮损症状识别[J]. 计算机应用, 2007,27(2): 492-493
14．陈增照 杨扬 何秀玲 喻莹 董才林 .基于核聚类的SVM多类分类方法[J]. 计算机应用, 2007,27(1): 47-49
15．薛志东 隋卫平 李利军.一种SVM与区域生长相结合的图像分割方法[J]. 计算机应用, 2007,27(2): 463-465
16．王强 陈英武 李孟军.一类支持向量机在烟叶选择中的应用[J]. 计算机应用, 2007,27(2): 482-485
17．周书仁;梁昔明;叶吉祥;朱灿.基于脸部信息和支持向量机的人脸检测[J]. 计算机应用, 2006,26(5): 1032-1034
18．蔺旭东 曾晓宁 薄静仪.一种基于支持向量的镜头聚类算法[J]. 计算机应用, 2007,27(9): 2143-2146
19．罗泽举 宋丽红 朱思铭.基于独立成分分析的分解向前SVM降维算法[J]. 计算机应用, 2007,27(9): 2249-2252
20．吕治国 徐昕 贺汉根.基于可变模板和支持向量机的人体检测[J]. 计算机应用, 2007,27(9): 2258-2261
21．曹晓莉 江朝元 甘思源.基于聚类支持向量机的船用污水处理装置故障诊断[J]. 计算机应用, 2008,28(10): 2648-2651
22．周红刚 杨春德 .基于免疫算法与支持向量机的异常检测方法[J]. 计算机应用, 2006,26(9): 2145-2147
23．柯永振 张家万 孙济洲 张怡 周小舟 .结合支持向量机与C均值聚类的图像分割[J]. 计算机应用, 2006,26(9): 2081-2083
24．崔江 王友仁 .基于聚类预处理和支持向量机的模拟电路故障诊断技术[J]. 计算机应用, 2006,26(8): 1977-1979
25．康恺 林坤辉 周昌乐 .基于主题词频数特征的文本主题划分[J]. 计算机应用, 2006,26(8): 1993-1995
26．衣杨 凌应标 常会友 肖志娇 .基于ε－SVR的销量预测规划计算模型和算法研究[J]. 计算机应用, 2006,26(8): 1968-1971
27．官金安 陈亚光 .通道选择对诱发脑电单次提取精度影响的研究[J]. 计算机应用, 2006,26(8): 1932-1934
28．李春茂 肖建 张玥.网络化控制系统两种时延预测算法及其比较[J]. 计算机应用, 2007,27(2): 257-260
29．黄聪 宣国荣 高建炯 施云庆 .基于图像及其预测误差图小波频域矩的隐写分析[J]. 计算机应用, 2006,26(8): 1851-1853
30．孔波; 刘小茂.基于中心距离比值的增量支持向量机[J]. 计算机应用, 2006,26(6): 1434-1436
31．李钢; 王蔚; 李乐加.支持向量机在脑电信号分类中的应用[J]. 计算机应用, 2006,26(6): 1431-1433
32．付长龙 吕彦波 姚全珠 杜旭辉.基于样本密度的SVM及其在入侵检测中的应用[J]. 计算机应用, 2007,27(4): 838-840
33．白裔峰 肖建 于龙 黄景春.基于结构风险最小化的加权偏最小二乘法[J]. 计算机应用, 2007,27(4): 939-941
34．沈新宇 许宏丽 官腾飞.基于直推式支持向量机的图像分类算法[J]. 计算机应用, 2007,27(6): 1463-1464
35．罗泽举 宋丽红 伍小明 詹希美.基于新型特征提取的寄生虫卵图像识别研究[J]. 计算机应用, 2007,27(6): 1485-1487
36．何振红 吕林涛.基于ICA-MJE和SVM的虹膜特征提取与识别[J]. 计算机应用, 2007,27(6): 1505-1507
37．李恒杰.Online SVM在实时入侵检测中的应用研究[J]. 计算机应用, 2007,27(6): 1339-1342
38．周辉仁 郑丕谔 赵春秀.基于遗传算法的LS-SVM参数优选及其在经济预测中的应用[J]. 计算机应用, 2007,27(6): 1418-1419
39．张秋余 刘洋.使用基于SVM的局部潜在语义索引进行文本分类[J]. 计算机应用, 2007,27(6): 1382-1384
40．王硕 周激流 彭博.基于API序列分析和支持向量机的未知病毒检测[J]. 计算机应用, 2007,27(8): 1942-1943
41．张慧档 贺昱曜.基于混沌序列的SVM参数选择及其在笔迹鉴别中的应用[J]. 计算机应用, 2007,27(8): 1961-

1963
42. 李爱媛 孟相如 张立 .基于SVM的故障诊断在网管平台中的应用[J]. 计算机应用, 2007,27(10): 2414-2416
43. 倪霖 郑洪英 .基于聚类和支持向量机的入侵检测研究[J]. 计算机应用, 2007,27(10): 2440-2442
44. 周书仁 梁昔明 杨秋芬 叶吉祥.基于PSO与ICA的表情特征提取[J]. 计算机应用, 2007,27(11): 2797-2799
45. 黄颖 李伟 刘发升.双隶属度模糊支持向量机算法[J]. 计算机应用, 2007,27(11): 2821-2824
46. 郭宇 孙敏.基于SVM成本决策分析模型的入侵响应研究[J]. 计算机应用, 2007,27(11): 2704-2706
47. 丰明聪 葛洪伟.基于可变区域特征和SVM的步态识别研究[J]. 计算机应用, 2007,(12): 3081-3083
48. 戴宏亮 戴道清.基于智能全间隔自适应模糊支持向量机的水质分类[J]. 计算机应用, 2008,28(11): 2847-2849
49. 何海江.代价与样本相关的简约核支持向量机[J]. 计算机应用, 2008,28(11): 2863-2866
50. 谢明霞 陈科 郭建忠.基于图谱理论的FCM图像分割方法研究[J]. 计算机应用, 2008,28(11): 2912-2914
51. 王金艳 冯建武 刘万里.一种不平衡支持向量机的校正方法[J]. 计算机应用, 2007,(12): 2896-2898
52. 吴广潮 闫丽 杨晓伟.基于模糊分割和邻近对的支持向量机分类器[J]. 计算机应用, 2008,28(1): 131-133
53. 刘陆洲 肖建.基于支持向量机的逆控制及其稳定性分析[J]. 计算机应用, 2008,28(11): 2978-2980
54. 张震 康吉全 平西建 任远.用统计特征量实现的图像拼接盲检测[J]. 计算机应用, 2008,28(12): 3108-3111
55. 张秋余 竭洋 李凯.基于模糊支持向量机与决策树的文本分类器[J]. 计算机应用, 2008,28(12): 3227-3230
56. 刘雪燕 李明 张亚芬.基于PCA和多约简SVM的多级说话人辨识[J]. 计算机应用, 2008,28(1): 127-130
57. 王东 吴湘滨.利用粒子群算法优化SVM分类器的超参数[J]. 计算机应用, 2008,28(1): 134-135,139
58. 王自强 钱旭.基于KDA和SVM的文档分类算法[J]. 计算机应用, 2009,29(2): 416-418
59. 徐锟 王万良 李祖欣.基于支持向量机的计算资源反馈调度[J]. 计算机应用, 2009,29(2): 535-538
60. 袁浩 付忠良 程建 阮波.基于支持向量机的纸张缺陷图像分类识别[J]. 计算机应用, 2008,28(2): 330-332,
61. 董建设 袁占亭 张秋余.基于多种核函数的SVM在垃圾邮件过滤中的应用[J]. 计算机应用, 2008,28(2): 424-427
62. 费玉莲 姜波 李渊.面向异步通讯机制的网页分类研究[J]. 计算机应用, 2008,28(2): 545-548
63. 吴宗亮 窦衡.一种广义最小二乘支持向量机算法及其应用[J]. 计算机应用, 2009,29(3): 877-879
64. 陈丽 陈静.基于支持向量机和k-近邻分类器的多特征融合方法[J]. 计算机应用, 2009,29(3): 833-835
65. 黄永文;何中市 伍星.用户评论的分类获取[J]. 计算机应用, 2009,29(3): 846-848
66. 张德贤 张苗 谭一鸣.基于启发式信息的支持向量机规则抽取[J]. 计算机应用, 2008,28(3): 729-731
67. 金展 范晶 陈峰 徐从富.基于朴素贝叶斯和支持向量机的自适应垃圾短信过滤系统[J]. 计算机应用, 2008,28(3): 714-718
68. 张宪 李晓娟.支持向量机在显微图像分类中的应用研究[J]. 计算机应用, 2008,28(3): 790-791
69. 解洪胜 张虹.基于内容的图像检索中SVM和Boosting方法集成应用[J]. 计算机应用, 2009,29(4): 979-981,
70. 刘明 王婷婷 黄小燕 刘锐.基于SVM分类区域的传感器网络节点自定位算法[J]. 计算机应用, 2009,29(4): 1064-1067
71. 傅丰 王端.一种改进的啤酒瓶分类识别技术[J]. 计算机应用, 2009,29(4): 1168-1170
72. 宋娇 葛临东.一种遗传模糊聚类算法及其应用[J]. 计算机应用, 2008,28(5): 1197-1199
73. 贝依林 闫德勤 梁宏霞 李克秋.基于支持向量机的彩色图像水印算法[J]. 计算机应用, 2008,28(5): 1247-1250
74. 肖建鹏 张来顺 任星.基于增量学习的直推式支持向量机算法[J]. 计算机应用, 2008,28(7): 1642-1644
75. 刘琼 周慧灿 王耀南.基于亮度分级和方向密度的无监督文本定位[J]. 计算机应用, 2008,28(6): 1523-1526
76. 张钊 费一楠 宋麟 王锁柱.基于模糊支持向量机的多分类算法研究[J]. 计算机应用, 2008,28(7): 1681-1683
77. 万明成 耿技 程红蓉 周俊怡.基于文本区域特征的图像型垃圾邮件过滤算法[J]. 计算机应用, 2008,28(8): 1904-1906
78. 蔡铁 伍星 李烨.集成学习中基于离散化方法的基分类器构造研究[J]. 计算机应用, 2008,28(8): 2091-2093
79. 吴石 耶夫戈尼耶·伊万诺维奇.基于小波特征和多类支持向量机的病态语音识别方法[J]. 计算机应用, 2008,28(8): 2097-2100
80. 陈蓉 宋俊德.基于SVM分块回归分析的话务量预测模型[J]. 计算机应用, 2008,28(9): 2230-2232
81. 姜贤林 郭秀清.基于支持向量机的质量控制软测量建模[J]. 计算机应用, 2008,28(9): 2382-2385
82. 许亮.融合先验知识的模糊最小二乘支持向量机模型及其应用[J]. 计算机应用, 2008,28(9): 2423-2426

83. 颜景斌 吴石 伊戈尔.艾杜阿尔达维奇.基于单类支持向量机的音频分类 [J]. 计算机应用, 2009,29(05): 1419-1422
84. 吴宗亮 窦衡.一种新的最小二乘支持向量机稀疏化算法 [J]. 计算机应用, 2009,29(06): 1559-1581
85. 卜令超 王士同.一种新的用于候选基因排序的数据融合方法 [J]. 计算机应用, 2009,29(06): 1563-1571
86. 许晓东 王传安 朱士瑞.基于信息熵SVM的ICMP隐蔽通道检测研究[J]. 计算机应用, 2009,29(07): 1796-1798
87. 王琳 闫德勤 梁宏霞.基于熵和蚁群聚类算法的模糊支持向量机[J]. 计算机应用, 2009,29(07): 1890-1893
88. 甘俊英 何思斌.基于2DLDA与SVM的人脸识别算法[J]. 计算机应用, 2009,29(07): 1927-1929
89. 李广明 刘群锋.光滑支持向量机两种求解算法的比较 [J]. 计算机应用, 2009,29(06): 1612-1614
90. 周欣然 滕召胜 赵新闻.基于LSSVM的MIMO系统快速在线辨识方法 [J]. 计算机应用, 2009,29(08): 2281-2284
91. 李广明 熊金志.光滑支持向量分类机的收敛上界研究 [J]. 计算机应用, 2009,29(08): 2243-2244
92. 许翔 张东波 黄辉先 刘子文.基于改进的粒子群算法和信息熵的知识获取方法 [J]. 计算机应用, 2009,29(08): 2245-2249
93. 王勇 刘九芬 张卫明.基于DCT系数多方向相关性的信息隐藏盲检测方法 [J]. 计算机应用, 2009,29(09): 2344-2347
94. 朱杰 李宁 高相辉.基于间隔聚类合并的支持向量机反问题求解算法[J]. 计算机应用, 2009,29(09): 2481-2482
95. 谢宏 刘敏 陈淑荣.基于ICA和SVM的道路网短时交通流量预测方法[J]. 计算机应用, 2009,29(09): 2550-2553
96. 张继宏 李小霞 孙波.基于非线性支持向量机的原核生物基因识别 [J]. 计算机应用, 2009,29(10): 2748-2750
97. 张艳秋 王蔚.利用遗传算法优化的支持向量机垃圾邮件分类 [J]. 计算机应用, 2009,29(10): 2755-2757
98. 赵冠华.基于二次Renyi熵的非迭代最小二乘支持向量机预测模型[J]. 计算机应用, 2009,29(10): 2751-2754
