

SVDD 参数优化的有限穷举—局部遗传算法及故障检测应用

蔡金燕, 杜敏杰, 孟亚峰, 朱 赛

(军械工程学院 光学与电子工程系, 河北 石家庄 050003)

摘 要: 针对支持向量数据描述(SVDD)训练过程中的参数优化问题, 提出了一种有限穷举—局部遗传算法。首先, 在分别分析参数 C 和 σ 对 SVDD 分类性能不同影响的基础上, 得到参数 σ 是影响分类性能主因的结论。然后针对 σ 的优化问题, 通过穷举有限个整数解并比较其分类性能来确定近似最优解, 在近似最优解的领域内用遗传算法进行局部搜索, 最终得到精确的优化参数。仿真实验及电路故障检测应用结果表明: 算法有效避免了参数搜索的盲目性, 能以更短的时耗逼近最优解。

关键词: 支持向量数据描述; 参数优化; 有限穷举—局部遗传算法; 故障诊断

中图分类号: TP26

文献标识码: A

文章编号: 1000-7180(2013)02-0013-05

A Parameter Optimization Method of Finite Exhaustion—Local GA for SVDD and Fault Diagnosis Application

CAI Jin-yan, DU Min-jie, MENG Ya-feng, ZHU Sai

(Optical and Electronic Engineering Department, Ordnance Engineering College, Shijiazhuang 050003, China)

Abstract: Aiming at the problem of SVDD parameter optimization in training process, an finite exhaustion—local genetic algorithm is proposed. Based on the analysis of C and σ parameters' different influence on SVDD classification performance respectively, a conclusion that σ is the main influencing factor is made. Focus on σ optimization issue, the near—optimum solution of σ is determined by enumerating finite integer solutions along with comparing these solutions' performance, and then search is launched locally in nearby domain of near—optimum solution. The precise parameter is get at last. Simulation experiment and circuit fault detection application results show that the above algorithm avoids blind parameter searching and can approach optimum solution in shorter time wastes.

Key words: SVDD; parameter optimization; finite exhaustion—local GA; fault diagnosis

1 引言

Tax 和 Duin^[1]提出的支持向量数据描述(Support Vector Data Description, SVDD)理论是当前流行的单分类器, 在图像检测、故障诊断等领域得到广泛应用。SVDD 核心思想是, 通过引入非线性变换和核函数, 将样本空间映射到特征空间, 在特征空间中寻求最小体积超球以尽可能多地包围目标样本, 因此该方法实质上是寻求样本的最优边界描述。在 SVDD 具体应用中, 折中参数 C 和核函数参数 σ 这

一对参数共同决定边界优劣, 因此如何确定最优参数对成为 SVDD 应用的关键环节。

网格搜索法是优化 SVDD 参数对的有效方法, 其本质是同时对两个参数的有限穷举。文献[1]提出参数 C 对分类性能影响有限, 可以按经验选取, 关键是参数 σ 的选取。文献[2]用交叉验证法对 σ 进行优化, 该方法时耗过大, 优化结果也不尽理想。文献[3]用粒子群优化(PSO)算法求解 σ , 优化结果具有随机性。文献[4]基于非高斯测度, 构造一个用于判别映射后的样本逼近超球形程度的准则函数, 使该函数值最小来

收稿日期: 2012-05-11; 修回日期: 2012-06-20

基金项目: 河北省重点基础研究项目(10963529D)

确定最优参数. 该方法需假设样本各分量服从独立同方差高斯分布. 本文针对 SVDD 参数优化问题, 分析了 C 和 σ 对分类器性能不同影响, 给出了 C 的经验范围, 提出了针对 σ 优化的有限穷举—局部遗传算法 (Genetic Algorithm, GA), 并将其应用到 SVDD 故障检测的参数优化问题中.

2 SVDD 基本理论

文献[1]在支持向量机的基础上, 第一次提出了 SVDD 理论, 可表述如下: 给定含 n 个样本的样本集 $X = \{x_i | x_i \in R^d, i = 1, 2, \dots, n, d \text{ 为样本维数}\}$, 通过非线性映射 $\varphi: R^d \rightarrow F$ 将样本映射到高维特征空间, 在 F 中寻求一个体积最小的超球体 $\Omega = (a, R)$, a 为球心 R 为半径, 使得样本集 $X = \{\Phi(x_i)\}_{i=1}^n$ 中的数据全部或尽可能多地包含在 Ω 中. 该问题是如式(1)的优化问题:

$$\min_{R, a, \xi_i} F(R, a, \xi_i) = R^2 + C \sum_{i=1}^n \xi_i \quad (1)$$

$$\text{s. t. } \|\varphi(x_i) - a\|^2 \leq R^2 + \xi_i, \xi_i \geq 0 \forall i$$

式中, C 为折中参数, 实现超球体积和错分样本比例间的折中, ξ_i 为松弛变量. 引入 Lagrange 乘子 $\alpha_i \geq 0, \gamma_i \geq 0$, 上述问题的解为如式(2)的 Lagrange 极值问题.

$$L(R, a, \alpha_i, \gamma_i, \xi_i) = R^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [R^2 + \xi_i - (\|\varphi(x_i)\|^2 - 2a \cdot \varphi(x_i) + \|a\|^2)] - \sum_{i=1}^n \gamma_i \xi_i \quad (2)$$

将 L 分别对 R, a, ξ_i 求偏导并用核函数 $K(x_i, x_j)$ (通常为高斯核函数) 代替内积 $\langle \varphi(x_i), \varphi(x_j) \rangle$ 可得如式(3)对偶问题.

$$\max L(\alpha_i) = \sum_{i=1}^n \alpha_i K(x_i, x_i) - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j)$$

$$\text{s. t. } \begin{cases} \sum_{i=1}^n \alpha_i = 1 \\ 0 \leq \alpha_i \leq C \end{cases} \quad (3)$$

任一测试样本 x 到超球球心的距离平方为:

$$D^2(x) = \|\varphi(x) - a\|^2$$

$$= K(x, x) - 2 \sum_{i=1}^n \alpha_i K(x, x_i) + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) \quad (4)$$

若 $D^2(x) \leq R^2$ 则判断 x 为目标类, 否则为异常类, 这里, 超球半径 R 通过超球边界上的任一支持向

量 x_w 到球心的距离计算得到. 上述就是 SVDD 基本理论.

3 SVDD 参数对分类性能影响分析

对于给定的训练样本集, 将单分类器 SVDD 应用于故障检测时, 分类性能受折中参数 C 和核函数参数 σ 影响很大. 最优的分类边界应同时兼顾检测率和误报率, 通常用接收者操作特性 (Receiver Operating Characteristics, ROC) 曲线及曲线下面积 Area Under an ROC Curve, AUC) 来表示这种分类性能.

3.1 C 对 AUC 的影响

在 SVDD 中, 通过参数 $C = 1/(T \times \text{fracrej})$ 来调控经验风险的大小, 实现超球体积和错分样本比例间的折中, 其中 T 为样本总数, fracrej 为正常样本的拒识比率, 用百分数表示. 故 $(T \times \text{fracrej})$ 的结果实际上决定了支持向量个数. 通常情况, 至少也有一个支持向量, 所以 $(T \times \text{fracrej})$ 的值至少是 1, 故 $1/T \leq C \leq 1$. 由此可见, C 的取值与支持向量个数紧密关联, 直接影响到 SVDD 的决策边界进而影响到分类器的 AUC. 为了更直观展示折中参数 C 对 AUC 的影响, 以下通过仿真调整 fracrej 的值 $1/T \leq \text{fracrej} \leq 1$ 来观察 AUC 的变化情况.

(1) 仿真数据为 100 个 2 维的 Banana 数据集另加 30 个随机生成的非目标数据集.

(2) 建模工具为 SVDD 工具箱 `dd_tools` 并选用高斯核函数.

(3) 仿真参数设定为: 核宽度固定为 $\sigma=5$, 拒绝率 fracrej 分别为 0.01、0.02、0.05、0.1、0.2、0.5.

(4) 仿真环境为: 酷睿双核 2.60GHz CPU, 2GB 内存, MATLAB2010.

图 1 为不同 fracrej 下样本在原空间的分类情况. 可见 fracrej 在一定程度上决定决策边界的伸缩, 不同大小的 fracrej 对应着不同的边界支持向量和非边界支持向量, 用这些分类结果对相同的测试样本将得到不同的分类性能. 同时, 从图中还可看出, 在 $0.01 \leq \text{fracrej} \leq 0.1$ 的范围内 ($1/T \leq \text{fracrej} \leq 10/T$), 决策边界的伸缩变化并不明显, 基本上能较为精确地包围目标样本; 只在 $\text{fracrej} \geq 0.1$ 时决策边界对样本描述才变得不理想.

为了更精确研究 fracrej 对分类性能的影响, 用 MATLAB 另生成 100 个 2 维的 Banana 数据集另加 30 个随机非目标数据集一并作为测试样本, 测试不同 fracrej 下的分类性能, 在 $0.01 \leq \text{fracrej} \leq 0.5$

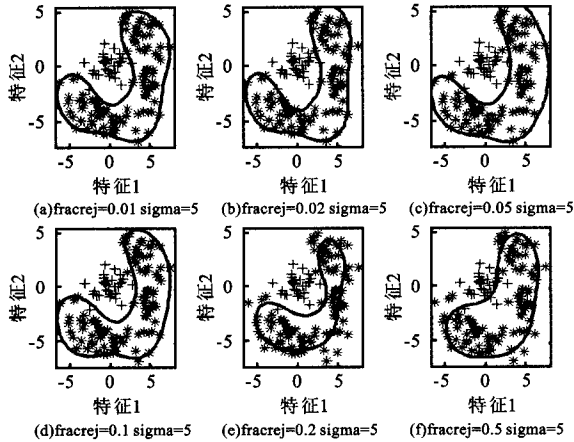


图 1 fracrej 对 SVDD 边界影响

内以 0.01 的步长改变 fracrej 的值,可以得到测试样本测试下 ROC 曲线,计算各曲线对应的 AUC,得到的 fracrej - AUC 关系如图 2 所示.可见,在 $0.01 \leq \text{fracrej} \leq 0.1$ 内, AUC 较理想且对 fracrej 变化不敏感.

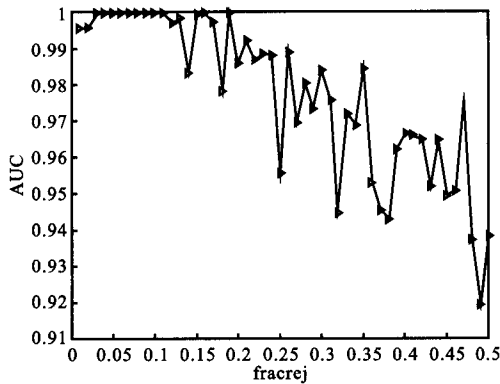


图 2 fracrej - AUC 关系

从以上分析仿真可得:SVDD 参数之一 C 的优化归根结底是对 fracrej 优化;在 $1/T \leq \text{fracrej} \leq 10/T$ 范围内,分类器保持着较高性能并且对 fracrej 变化不敏感.因此在参数优化过程中,fracrej 可直接按经验取值,一般取 $\text{fracrej} = 5/T$.

3.2 核参数 σ 对 AUC 的影响

前已提及,核函数实现了输入空间到特征空间的映射,因此,若确定了核函数的具体形式,就相当于确定了映射函数集对应的特征空间,这样核函数参数 σ 的改变实际上是隐含地改变对应的映射函数,从而改变输入空间分布的复杂程度.为了直观展示核函数参数 σ 对 AUC 的影响,以下通过仿真调整 σ 的值来观察 AUC 的变化情况.

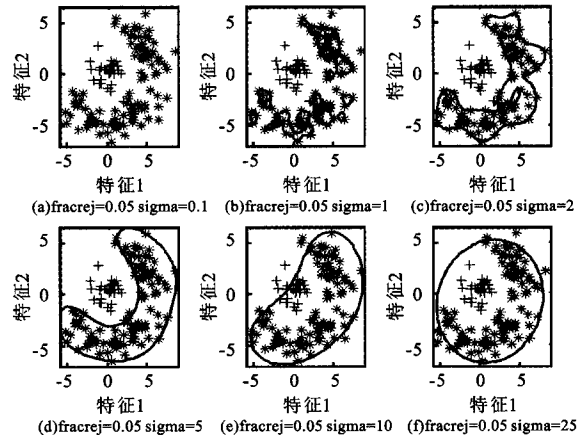
(1) 仿真数据为 100 个 2 维的 Banana 数据集另加 30 个随机生成的非目标数据集.

(2) 仿真参数设定为拒绝率 $\text{fracrej} = 0.05$,核函

数参数 σ 分别为 0.1、1、2、5、10、25.

(3) 建模工具、仿真环境不变.

图 3 为不同 σ 下样本在输入空间的分类情况.可见随着 σ 的增大分类器逐渐由过学习变为欠学习,不同大小的 σ 对决策边界形状有明显影响,用这些分类结果对相同的测试样本将得到不同的分类性能.事实上, σ 极小时,相当于每个目标样本都单独聚为一类; σ 极大时,超球无限大,只有一个支持向量.

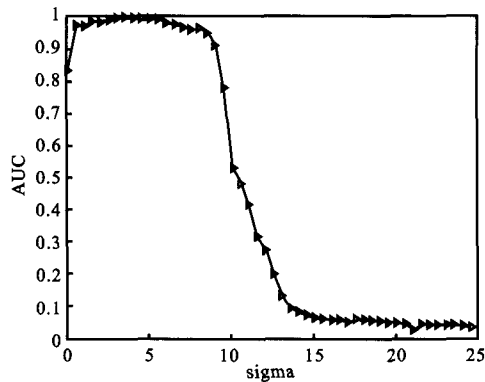
图 3 σ 对 SVDD 边界影响

为了检验不同 σ 下的分类性能,用 MATLAB 另生成 100 个 2 维的 Banana 数据集另加 30 个随机非目标数据集一并作为测试样本,测试不同 σ 下的训练效果,图 4 给出了训练样本测试下的 σ - AUC 关系.可见,分类器性能受 σ 影响波动较大: $0 < \sigma < 10$ 时分类器 AUC 保持在较高水平在该范围内存在某个最优的 σ ,使得该参数训练下的分类器 AUC 最大; $\sigma > 10$ 时,分类器性能急剧下降.

综上所述,有以下结论:拒绝率 fracrej 决定 SVDD 决策边界的伸缩,在 $1/T \leq \text{fracrej} \leq 10/T$ 范围内分类器性能较好且对 fracrej 变化不敏感;核函数参数 σ 决定决策边界的形状,进而决定过学习或欠学习.因此,对参数对 $(\text{fracrej}, \sigma)$ 处理的方法是:fracrej 按经验取, σ 通过智能优化的方法求得.

3.3 参数优化的有限穷举—局部 GA

通过以上分析知:SVDD 参数优化的关键是对 σ 的优化,本文针对该问题,提出了一种基于有限穷举—局部 GA 算法的 SVDD 参数 σ 优化方法.基本思想是:首先,将 σ 限定在可能的区域内,在该区域内取有限个整数点作为训练 SVDD 所需的 σ 参数,计算各自的 AUC,记录具有最大 AUC 的整数 σ ;然后,在 σ 的局部小范围内用 GA 搜索最优解.穷举阶段避免了群智能算法搜索的盲目性;局部寻优阶段选择 GA 作为智能搜索工具,GA 能够使解决的问

图4 σ -AUC 关系

题逐步逼近最优或近优解。通过该算法,有效克服了 PSO 等算法优化过程中速度慢和结果不理想的问题,其流程如图 5 所示。

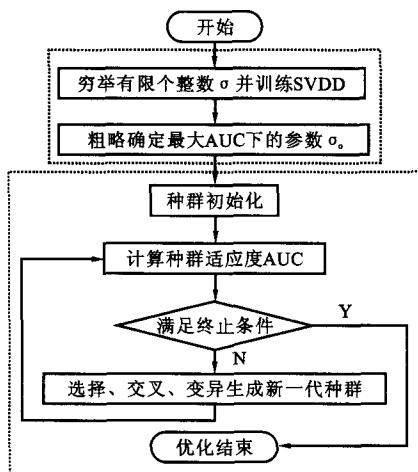


图5 有限穷举—局部 GA 算法流程

图 5 中小的虚线框表示有限穷举过程,大的虚线框表示局部 GA 过程。以下给出算法具体步骤。

Step 1. 从整数 1 开始,穷举有限个整数 σ 并用其训练 SVDD。

Step 2. 计算不同 σ 下的 AUC,记录近似最优解 $\sigma_0 = \arg\max(\text{AUC})$ 。

Step 3. 在 σ_0 的领域内对变量 σ 进行基因编码,生成一定规模的染色体构成种群,随机初始化种群。

Step 4. 采用轮盘赌策略确定种群内个体适应度(适应度函数为 AUC)并判断是否满足终止条件,若满足则输出最优个体并结束计算,否则转下一步。

Step 5. 根据适应度对种群内的个体进行选择、交叉、变异生成新一代种群并返回 Step2。

4 实验验证与分析

为检验算法有效性,用 2 维的 Banana 数据集进行验证,生成 100 个 Banana 数据集和 30 个随机样本一并作为训练样本,再生成 100 个目标样本和 30

个随机样本一并作为测试样本。使用 SVDD 工具箱 dd_tools 建模并选用高斯核函数,GA 种群规模为 10 个,进化迭代次数为 20,交叉概率 0.9,变异概率 0.05。仿真环境不变。优化结果为最优的 SVDD 参数为 3.9,对应的 AUC 为 0.998,CPU 时耗 183s。

为检验算法通用性,任从 UCI^[5] 的机器学习库中选取若干真实数据集验证。具体如下:

(1) Ionosphere. 训练样本:110 个 good 类目标样本外加 60 个 bad 类非目标样本;测试样本:115 个 good 类目标样本外加 66 个 bad 类非目标样本。

(2) Breastcancer. 训练样本:100 个 benign 类目标样本外加 50 个 malignant 类非目标样本;测试样本:300 个 benign 类目标样本外加 200 个 malignant 类非目标样本。

(3) Vehicle. 训练样本:150 个 van 类目标样本外加 50 个 bus 类非目标样本;测试样本:50 个 van 类目标样本外加 20 个 bus 类非目标样本。

参数均设定同上,仿真环境不变。表 1 列出了参数 σ 优化结果及其与 PSO 优化性能比较 ($\text{fracrej} = 0.05$)。

表1 UCI 数据集优化结果

UCI 数据	σ	AUC	本文 时耗(s)	PSO 时耗(s)
Ionosphere	3.5	0.98	278	526
Breastcancer	0.85	0.99	182	348
Vehicle	1.55	0.96	374	714

可以看到,本节提出的有限穷举—局部 GA 有效优化了 SVDD 的训练参数,在同等优化结果下较之 PSO 算法费时更少。

为验证算法在具体应用中的优化效果,以图 6 所示的视频放大电路^[6]为例进行验证。电路激励为 5V、100Hz 正弦电压源信号,电阻、电容、电感元件容差均设定为标称值的 1%,故障设置为: F_1 故障 ($C_1 + 50\%$)、 F_2 故障 ($C_2 + 50\%$)、 F_3 故障 ($C_3 + 50\%$)、 F_4 故障 ($R_1 + 50\%$)、 F_5 故障 ($R_2 + 50\%$)、 F_6 故障 ($R_3 + 50\%$)、 F_7 故障 ($R_4 + 50\%$)、 F_8 故障 ($R_5 + 50\%$)、 F_9 故障 ($R_6 + 50\%$),用蒙特卡罗分析法仿真故障数据,分别采集测试点 N_1 、 N_2 、 N_3 、 N_4 上的电压信号作为特征向量,用 100 个正常样本作为目标样本、45 个各类故障样本作为异常样本一并作为训练样本,训练参数 fracrej 取 0.05,用文中的有限穷举—局部 GA 优化 σ ,得到最优的 σ 值为 1.2。用最优的参数训练 SVDD,并用另外 50 个正常样本作为目标样本、25 个故障样本作为异常样本一并作为测试

样本,测试得到最优参数下的分类器 AUC 为 0.92,耗时 221s,而用 PSO 算法在同样耗时下的最大 AUC 仅为 0.77.

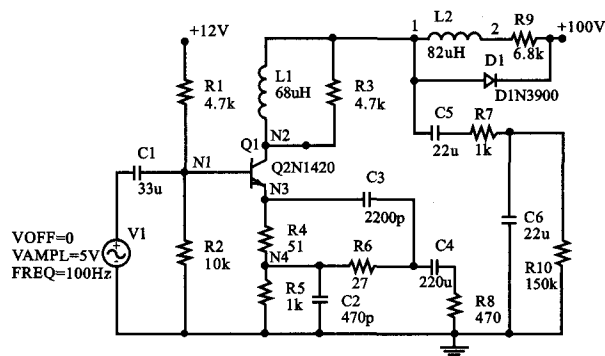


图6 视频放大电路原理图

5 结束语

文中针对 SVDD 训练过程中的参数优化问题,提出了一种有限穷举—局部遗传算法,并将其应用到电路故障检测中.仿真验证了算法有效性,电路故障检测应用表明:算法能以更短的耗时逼近最优解.

参考文献:

- [1] Tax D, Duin R. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20(11/13): 1191—1199.
- [2] 徐图,何大可. 超球体多类支持向量机理论[J]. 控制理论与应用, 2009, 26(11): 1293—1297.
- [3] 唐明珠,王岳斌,阳春华. 一种改进的支持向量数据描述故障诊断方法[J]. 控制与决策, 2011, 26(7): 967—972.
- [4] 赵峰,张军英,刘敬. 一种改善支撑向量域描述性能的核优化算法[J]. 自动化学报, 2008, 34(9): 1122—1127.
- [5] Newman D J, Hettich S, Blake C L, et al. UCI repository of machine learning databases[EB/OL]. (2008—10—19). <http://www.ics.uci.edu/~mllearn/ML-Repository.html>.
- [6] 汪鹏,杨士元. 模拟电路故障诊断测试节点优选新算法[J]. 计算机学报, 2006, 29(10): 1780—1785.

作者简介:

蔡金燕 女,(1961—),教授,博士生导师.研究方向为检测理论、故障诊断、可靠性共性技术.

杜敏杰 男,(1984—),博士研究生.研究方向为武器系统性能检测与故障诊断.

孟亚峰 男,(1971—),博士,副教授.研究方向为武器系统性能检测与故障诊断.

朱赛 男,(1987—),博士研究生.研究方向为武器系统性能检测与故障诊断.

(上接第12页)

- [6] 赵仕俊,孙美玲,唐懿芳. 基于遗传模拟退火算法的无线传感器网络定位算法[J]. 计算机应用与软件, 2009, 26(10): 189—192.
- [7] 张原. 基于高斯混合模型的无线传感器网络节点定位算法的研究[D]. 吉林,吉林大学, 2010.
- [8] Jang-Ping Sheu, Jian-Ming Li. A distributed location estimating algorithm for wireless sensor networks [C] // Bob Werner. Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing. Taichung: IEEE Press, 2006.
- [9] 朱剑,赵海,徐久强,等. 无线传感器网络中的定位模型[J]. 软件学报, 2011, 22(7): 1612—1625.
- [10] Neal Patwari, Alfred O, Hero III, matt perkins, neiyer correal, and robert [EB/OL]. [2006—10—17]. <http://www.eecs.umich.edu/~hero/localize/>.

作者简介:

叶苗 男,(1977—),博士研究生.副教授.研究方向为网络技术.进化计算.

程小辉 男,(1961—),博士,教授.研究方向为计算机网络技术、嵌入式系统、数据库技术.

邓昀 男,(1980—),硕士,讲师.研究方向为嵌入式系统、计算机网络.