第35卷 第22期 计 算 机 工 程 2009年11月
Vol.35 No.22 Computer Engineering November 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)22—0210—02 文献标识码：A 中图分类号：TP181

改进的SVDD增量学习算法

花小朋，皋 军，田 明，刘其明
(盐城工学院信息工程学院，盐城 224001)

摘 要：通过对SVDD增量学习中原样本和新增样本的特性分析，提出一种改进的SVDD增量学习算法。在增量学习过程中，该算法选取原样本的支持向量集和非支持向量中可能转为支持向量的样本集以及新增样本中违反KKT条件的样本作为训练样本集，舍弃对最终分类无用的样本。实验结果表明，该算法在保证分类精度的同时减少了训练时间。

关键词：支持向量数据描述；KKT条件；支持向量；增量学习

Improved Incremental Learning Algorithm for Support Vector Data Description

HUA Xiao-peng, GAO Jun, TIAN Ming, LIU Qi-ming
(School of Information Engineering, Yancheng Institute of Technology, Yancheng 224001)

【Abstract】 An improved incremental learning algorithm for Support Vector Data Description(SVDD) is presented through the characteristic analysis of old samples and new samples. In the course of incremental learning, support vecter set and non-support vector set which may be converted into support vector in old samples and samples which violate Karush-Kuhn-Tucker(KKT) condition in new samples are chosen as training samples and the useless samples are discarded in this algorithm. Experimental results show that the training time is greatly reduced while the classification precision is guaranteed.

【Key words】 Support Vector Data Description(SVDD); Karush-Kuhn-Tucker(KKT) condition; support vector; incremental learning

1 概述

支持向量数据描述(Support Vector Data Description, SVDD)是由文献[1]提出发展起来的一种单值分类算法，其理论源于文献[2]提出的支持向量机。

在说话人识别[3]、入侵检测[4]、机械故障诊断[5]、气象预报[6]等领域的应用中，SVDD算法均有很好的识别效果。但在实际应用中发现，当训练样本集过于庞大、无法一次性读入内存时，就需要把训练集分成几个独立的子集，依次在各个子集上作增量学习；另外，要在训练初期就收集一个非常完整的训练集是非常困难甚至是难以实现的，而在更多的情况下，样本是不断加入的，即增量式地加入训练样本，因此希望SVDD具有这样的能力，即其学习的精度可以随着应用过程中样本集的不断积累而逐步提高。

已有先进的SVDD训练方法如SMO[7]、可行方向法[8]等本身都无法进行增量学习，因此研究SVDD的增量学习方法是有意义的。SVDD的学习结果为支持向量集，通常是整个样本集的一小部分，但能描述整个样本集的分类特征，这说明研究基于SVDD的增量学习方法是可行的。

基于支持向量集的常规增量学习算法[9]是将旧样本集的支持向量加入到新增样本中进行训练学习，从而得到新的分类模型。这种算法存在2个缺陷：

(1)随着新增样本的加入，旧样本集中非支持向量可能会转换成新的支持向量，过快的样本丢弃率会影响增量学习器的分类精度；

(2)新增样本并不一定是全值得学习的，对于旧样本集能体现的新增样本无需重复学习。

基于此分析，本文提出了一种改进的SVDD增量学习算法IISVDD。

2 支持向量数据描述理论

2.1 支持向量数据描述(SVDD)

根据文献[1]，SVDD算法的思想为寻找一个超球体，在使其半径尽可能小的同时，包含的训练样本数尽可能多。其目标函数为

$$\min_R R^2 + C\sum_i \xi_i \tag{1}$$

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

其中，R为球体半径；a为球心；ξ为松弛变量；C为正则化参数，控制对错分样本的惩罚程度，实现对球的大小和所包含的样本数之间的折衷。同SVM类似，当样本点非线性可分时，使用非线性映射$\Phi(x_i)$，将样本点映射到高维特征空间，在那里样本是线性可分的，并选择满足Mercer条件的核函数$K(x,y)=(\Phi(x),\Phi(y))$，其中，$(\Phi(x),\Phi(y))$表示$\Phi(x)$与$\Phi(y)$的内积。

式(1)和式(2)的对偶问题为

$$W(\alpha) = \min_\alpha \sum_i \sum_j \alpha_i \alpha_j K(x_i, x_j) - \sum_i \alpha_i K(x_i, x_j) \tag{3}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, \sum_i \alpha_i = 1 \tag{4}$$

基金项目：盐城工学院重点学科建设基金资助项目(XKY2007065)
作者简介：花小朋(1975－)，男，讲师、硕士，主研领域：人工智能，识别技术；皋 军，副教授、博士；田 明、刘其明，副教授、硕士
收稿日期：2009-09-15 **E-mail：**xp_hua@163.com

其中，α_i 为拉格朗日乘子，求解上述对偶问题可得最优解 α_i^*。事实上只有对应的少数样本点称为支持向量，体现在超球体的球面上(包括边界)。超球体的半径由边界上的任一支持向量点到球心的距离决定。一个测试点 z 被接受为目标样本，只需满足：

$$f(z)=K(z,z)-2\sum_i \alpha_i K(z,x_i)+\sum_{i,j}\alpha_i\alpha_j K(x_i,x_j)\le R^2 \tag{5}$$

2.2 KKT 条件

根据最优化原理，当所有 α_i 满足目标函数的 Karush-Kuhn-Tucker(KKT)条件，就可认为是原方程的一个解。文献[7]通过推导给出了目标函数的 KKT 条件：

$$\alpha_i=0\Rightarrow d_i^2\le R^2 \tag{6}$$

$$0\prec\alpha_i\prec C\Rightarrow d_i^2=R^2 \tag{7}$$

$$\alpha_i=C\Rightarrow d_i^2\ge R^2 \tag{8}$$

其中，d_i^2 为样本点 x_i 到球心 a 的距离的平方。

3 SVDD 增量学习过程理论分析

定理 1 在分类 SVDD 中，对应于 $\alpha_i=0$ 的样本分布于超球内(包含超球边界)；对应于 $0\prec\alpha_i\prec C$ 的样本分布于超球边界上；对应于 $\alpha_i=C$ 的样本分布于超球外(包含超球边界)。

证明：由上述 KKT 条件即可得证。

在 SVDD 增量学习中，由于新增训练样本并未得到原 SVDD 的学习，因此令此时这些样本所对应的 Lagrange 乘子 α_i 皆为 0。

3.1 新增样本理论分析

定理 2 若新增样本均满足 KKT 条件，则新增样本中肯定不存在新支持向量(新支持向量是指原样本和新增样本合并训练后的支持向量)。

证明：新增样本的 Lagrange 乘子皆为 0，由于新增样本满足原 SVDD 的 KKT 条件，则由定理 1 可知，新增样本分布于超球内(包含超球边界)，这样 SVDD 对合并后的样本进行优化与对原样本优化结果是等价的，新增样本的 Lagrange 乘子(皆为 0)与原样本的 Lagrange 乘子组合后生成的新 Lagrange 乘子就是新 SVDD 的最优解。由于新增样本的 Lagrange 乘子皆为 0，因此新增样本中不存在新支持向量。

定理 2 得证。

定理 3 若新增样本存在违背 KKT 条件的样本，则违背 KKT 条件的样本中必存在新支持向量。

反证法：

证明：由于新增样本中违背 KKT 条件的样本中不存在新支持向量，结合定理 2 可进一步推知，新增样本中不存在新支持向量。因此，与原样本合并训练后，新增样本的 Lagrange 乘子皆为 0，这样 SVDD 对合并后的样本进行优化与对原样本优化结果是等价的。由于新增样本肯定满足合并优化后 SVDD 的 KKT 条件，因此也应该满足原 SVDD 的 KKT 条件。显然与已知条件矛盾。

定理 3 得证。

由定理 2、定理 3 可推知，对于新增样中满足 KKT 条件的样本，由于原样本集已经包含了这部分样本的信息，因此无需再对这些样本进行学习；对于新增训练样本中那些违反 KKT 条件的样本，说明原样本集中没有包含这部分样本的信息，需要对这部分样本进行学习。

3.2 原样本理论分析

定理 4 若新增样本中存在违背 KKT 条件的样本，则原样本中非支持向量可能会转为支持向量(因篇幅有限，此定理可参照文献[10]中定理 3 通过实例进行证明)。

由定理 4 可推知，原样本集中非 SV 数据在增量学习过程中不能随便舍弃，因为这些非 SV 数据中有部分数据可能在增量学习后又转化成新的 SV。实验中发现，这些可能转换为新 SV 的数据通常位于超球面附近，数学表示为

$$R-\theta\le f(z)\le R,\theta\in[0,R]$$

这里 θ 与样本分布有关，样本分布越松散，θ 越大；同时当样本的统计性质比较差时，原样本和新增样本分布不相似，也会导致 θ 值偏大。另外，随着增量学习过程的增加，SV 附近的样本点越来越多，相应地 θ 越来越小，则原样本中非 SV 转化为 SV 的可能性越来越小。至于新训练样本加入后，原 SV 集的变化可以不予考虑，只要在增量学习中加入原 SV 集即可，因为原 SV 集的数据量极少，不会影响增量学习速度。

4 一种改进的 SVDD 增量学习算法 IISVDD

通过上述增量学习过程的理论分析，本文提出一种改进的 SVDD 增量学习算法 IISVDD。增量学习过程中新训练集的组成为原 SV 集、原非 SV 集中满足 $R-\theta\le f(z)\le R,\ \theta\in[0,R]$ 条件的样本以及新增样本中违反 KKT 条件的样本。具体算法描述如下：

输入：原始样本集 X_O，新增样本集 X_I。

(1)用 X_O 训练得到 SVDD 分类器 Ω_O，同时得到支持向量集 SV_O，非支持向量集 NSV_O。

(2)检验 X_I 中是否存在违反 Ω_O 的 KKT 条件的样本，若没有，算法终止，Ω_O 为学习结果；否则，选取其中违反 KKT 条件的样本 $X_{I\text{-}V}$，删除其余样本。

(3)选取 NSV_O 中满足 $R-\theta\le f(z)\le R,\ \theta\in[0,R]$ 条件的样本 $NSV_{O\text{-}S}$，删除其余样本。

(4)将 SV_O, $X_{I\text{-}V}$ 和 $NSV_{O\text{-}S}$ 合并训练得到最终分类器 Ω。

输出：增量学习后的分类器 Ω。

5 实验比较分析

实验数据采用 UCI 数据库中的 Optical Recognition of Handwritten Digits(ORHD)数据集[11]。该数据集共包括 3 823 个训练数据和 1 797 个测试数据，分为 10 类，每条数据为 64 维向量，笔者选择 0~4 类数据进行实验分析(见表 1)。实验环境为 Intel 迅驰 1.4 GHz、256 MB 内存、Win XP 操作系统和 Matlab 6 工具软件。

表 1 ORHD 数据集(0~4 类)的构成

数据类型	0 类	1 类	2 类	3 类	4 类
训练集	376	389	380	389	387
测试集	178	182	177	183	181

实验设计：对每类数据单独实验，初始样本来自各自训练集，数量分别为 76(0 类)、89(1 类)、80(2 类)、89(3 类)和 87(4 类)；增量样本来自各类训练集中其余样本，都分三步取，分别为 $C1$=50, $C2$=100, $C3$=150。对各类训练集的学习结果用各自的测试集检验。实验中核函数均采用高斯核，即 $K(x,y)=\exp(-\|x-y\|^2/2\sigma^2)$, C =0.1, σ =50, θ =0.05 R。为了验证文中所提增量学习算法 IISVDD 的性能，笔者将其与常规增量学习算法及非增量学习算法进行了实验比较，实验结果如表 2 所示。从表 2 的实验结果可推知，文中所提增量学习算法 IISVDD 很大程度上简化了训练集，从而使训练时间极大地减少，同时分类正确率得到了保障。

(下转第 215 页)