第35卷 第21期 计 算 机 工 程 2009年11月
Vol.35 No.21 Computer Engineering November 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)21—0183—03 文献标识码：A 中图分类号：TP181

基于简单特征值问题的修正 GEPSVM

徐金宝[1]，业巧林[2]，业 宁[2]
(1. 南京工程学院计算机工程学院，南京 211167；2. 南京林业大学信息技术学院，南京 210037)

摘 要：针对 GEPSVM 可能存在奇异性问题，提出一个改进的 GEPSVM 算法——SMGEPSVM。基于简单特征值问题，得到 2 个非平行过原点的超平面。与 GEPSVM 相比，SMGEPSVM 不仅可以求解 XOR 问题，还可求解简单特征值问题，避免 GEPSVM 奇异性问题，测试精度好于 GEPSVM，学习速度更快，其计算结果在人工数据集与 UCI 上得到了验证。

关键词：XOR 问题；奇异性问题；广义特征值；简单特征值

Modified GEPSVM Based on Simple Eigen-values Problems

XU Jin-bao[1], YE Qiao-lin[2], YE Ning[2]
(1. School of Computer Engineering, Nanjing Institute of Technology, Nanjing 211167;
2. School of Information Technology, Nanjing Forestry University, Nanjing 210037)

【Abstract】 Aiming at the flaw of singular problems appearing in GEPSVM, this paper proposes an improved GEPSVM——SMGEPSVM. Two non-parallel surfaces are obtained by solving simple Eigen-values problems. Compared with GEPSVM, the approach can solve XOR problems, and it further solves the simple Eigen-value problems, avoids the singular problems appearing in GEPSVM, it has better test set correctness and faster training time. Computational results on artificial datasets and UCI indicate the better performance of SMGEPSVM.

【Key words】 XOR problems; singular problems; generalized Eigen-values; simple Eigen-values

1 概述

支持向量机(Support Vector Machine, SVM)方法建立在统计学习理论中的结构风险最小化和 VC 维基础上[1-2]。传统的 SVM 分类技术通过求解凸规划问题来获得最优分类面[3]，其时间复杂度为 $O(n^{3.5})$ [4]。

文献[5]提出的最接近支持向量机(PSVM)用等式约束替换传统 SVM 不等式约束，使求解 SVM 凸规划问题转变为对线性方程组的求解。其时间复杂度为 $O(n^3)$ [6]。PSVM 的本质在于保持 2 类样本间间隔尽可能大，用 2 个平行超平面来拟合 2 类样本。文献[6]提出一种推广型 PSVM——GEPSVM。该算法摒弃了 PSVM 两平面平行的约束，主要通过求解广义特征值问题下最小特征值对应的特征向量来获得 2 个不平行的分类面。每个分类面要求离本类样本点尽可能近，而离其他类别的样本尽可能远。它能够有效地解决 XOR 问题。当解决 XOR 问题时，线性 GEPSVM 的分类精度达到了 100%，而 SVM 只能获得 46%的分类精度。尽管 GEPSVM 获得了成功，但其存在着很多缺陷：(1)可能存在奇异性问题；(2)从文献[1]的实验结果中可进一步看出，GEPSVM 在真实数据集中的分类性能稍劣于 SVM；(3)其学习速度远不及 PSVM[2]。针对 GEPSVM 以上的缺陷，本文提出一个基于简单特征值问题的修正多平面最接近支持向量机——SMGEPSVM。

2 GEPSVM 简介

在 n 维输入空间 R^n 上，给定 m 个训练样本，其由 $m_1\times n$ 的表示正类样本集的矩阵 $\boldsymbol{A}$ 与 $m_2\times n$ 的表示负类样本集的矩阵 $\boldsymbol{B}$ 定义，这里 $m_2+m_1=m$。

GEPSVM 主要的目的是在 n 维输入空间上输入空间上找到 2 个最优不平行的超平面，即：

$$\boldsymbol{x}'\boldsymbol{w}^1-\boldsymbol{e}^1r^1=0,\ \boldsymbol{x}'\boldsymbol{w}^2-\boldsymbol{e}^2r^2=0 \tag{1}$$

其要求每一个分类面离本类样本点尽可能近，而离其他类别的样本尽可能远。

二分类 GEPSVM 分类器能够被写成如下优化形式：

$$\min_{(\boldsymbol{w}^1,r^1)\neq 0}\frac{\|\boldsymbol{A}\boldsymbol{w}^1-\boldsymbol{e}^1r^1\|^2+\delta\left\|\begin{bmatrix}\boldsymbol{w}^1\\ r^1\end{bmatrix}\right\|^2}{\|\boldsymbol{B}\boldsymbol{w}^1-\boldsymbol{e}^1r^1\|^2} \tag{2}$$

$$\min_{(\boldsymbol{w}^2,r^2)\neq 0}\frac{\|\boldsymbol{B}\boldsymbol{w}^2-\boldsymbol{e}^2r^2\|^2+\delta\left\|\begin{bmatrix}\boldsymbol{w}^2\\ r^2\end{bmatrix}\right\|^2}{\|\boldsymbol{A}\boldsymbol{w}^2-\boldsymbol{e}^2r^2\|^2} \tag{3}$$

其中，$\boldsymbol{e}^i=[1,1,\cdots,1]'_{m_i\times 1}, i\in 1,2$。

式(2)的优化准则即在最小化目标下，分子要尽可能小，分母尽可能大，分子为第 1 类样本到第 1 个超平面距离的平方和，而分母表示第 2 类样本到第 1 个超平面距离的平方和，即要第 1 类超平面离第 1 类样本距离尽可能近，而距第 2 类样本距离尽可能远。记为

$$\boldsymbol{G}=[\boldsymbol{A}\ -\boldsymbol{e}^1]^{\mathrm{T}}[\boldsymbol{A}\ -\boldsymbol{e}^1]+\delta\boldsymbol{I},\ \boldsymbol{H}=[\boldsymbol{B}\ -\boldsymbol{e}^1]^{\mathrm{T}}[\boldsymbol{B}\ -\boldsymbol{e}^1]$$
$$\boldsymbol{L}=[\boldsymbol{B}\ -\boldsymbol{e}^2]^{\mathrm{T}}[\boldsymbol{B}\ -\boldsymbol{e}^2]+\delta\boldsymbol{I},\ \boldsymbol{M}=[\boldsymbol{A}\ -\boldsymbol{e}^2]^{\mathrm{T}}[\boldsymbol{A}\ -\boldsymbol{e}^2]$$

基金项目：国家自然科学基金资助项目(30671639)；2008 年江苏省研究生科技创新基金资助项目；南京工程学院科研基金资助项目(KXJ08071)

作者简介：徐金宝(1970－)，男，讲师、硕士，主研方向：Java 新技术与 MIS 研制，数据挖掘；业巧林，硕士；业 宁，副教授、博士

收稿日期：2009-04-16 **E-mail：**keeboo@163.com

$$z^1=\begin{bmatrix} w^1 \\ r^1 \end{bmatrix},\quad z^2=\begin{bmatrix} w^2 \\ r^2 \end{bmatrix}$$

则式(2)、式(3)能被式(4)重写，即：

$$\min_{z^1\neq 0}\frac{z^1{}'Gz^1}{z^1{}'Hz^1},\ \min_{z^2\neq 0}\frac{z^2{}'Lz^2}{z^2{}'Mz^2} \tag{4}$$

很明显，式(4)是 rayleigh 商问题，因此，可以用其相应的有用的性质来求解它，即求解 2 个广义特征值问题：$Gz^1=\lambda^1 Hz^1$ 和 $Lz^2=\lambda^2 Mz^2$。

3 线性 SMGEPSVM

SMGPSVM 的优化目标和 GEPSVM 相同，本文的思想是产生 2 个不平行的超平面，并保证每个超平面离本类样本尽可能近，离它类样本尽可能地远。为了简单起见，只先考虑过原点的 2 个平面。同时为了避免 GEPSVM 可能出现的奇异性问题，对式(4)做如下修正：

$$\begin{aligned}&\min\ \sum_{i=1}^{n_1}(w_1{}^t x_i^1)^2-\mu\sum_{i=1}^{n_2}(w_1{}^t x_i^2)^2\\&\text{s.t}\ \|w_1\|=1\end{aligned} \tag{5}$$

$$\begin{aligned}&\max\ \sum_{i=1}^{n_1}(w_2{}^t x_i^1)^2-\mu\sum_{i=1}^{n_2}(w_2{}^t x_i^2)^2\\&\text{s.t}\ \|w_2\|=1\end{aligned} \tag{6}$$

其中，μ 是正常数，是优化平衡因子；n_1 和 n_2 分别指第 1 类样本与第 2 类样本的个数。

先来求解使式(5)达到最小值的最优方向 w_1。式(5)能被重写为

$$\begin{aligned}&\min\ w_1{}'(S_1-\mu S_2)w_1\\&\text{s.t}\ \|w_1\|=1\end{aligned} \tag{7}$$

其中，$S_1=\sum_{i=1}^{n_1}x_i^1x_i^1{}'$; $S_2=\sum_{i=1}^{n_2}x_i^2x_i^2{}'$。

定理 1 第 1 类原型超平面的最优方向 w_1 是矩阵 $(S_1-\mu S_2)$ 最小特征值对应的特征向量。

证明 引入 Lagrange 函数

$$L(w_1,\lambda_1)=\sum_{i=1}^{n_1}(w_1{}^t x_i^1)^2-\mu\sum_{i=1}^{n_2}(w_1{}^t x_i^2)^2-\lambda_1(w_1{}^t w^1-1) \tag{8}$$

令 $\frac{\partial L}{\partial w_1}=0$ 整理得

$$\left(\sum_{i=1}^{n_1}x_i^1x_i^1{}'-\mu\sum_{i=1}^{n_2}x_i^2x_i^2{}'\right)w_1=\lambda^1 w_1 \tag{9}$$

即

$$(S_1-\mu S_2)w_1=\lambda^1 w_1 \tag{10}$$

将式(10)代入式(7)中，有

$$\min\ w_1{}'(S_1-\mu S_2)w_1=\min\ \lambda_1 w_1{}'w_1=\min\ \lambda_1$$

因此，第 1 类原型超平面的最优方向 w_1 是矩阵 $(S_1-\mu S_2)$ 最小特征值对应的特征向量。

记 w_1 *为由式(5)确定的最优方向，r_1 为第 1 类原型超平面的阈值。设：

$$r_1=\frac{1}{n_1}\sum_{i=1}^{n_1}w_1{}^*{}'x_i^1 \tag{11}$$

这样，确定的第 1 类超平面为

$$w_1{}^*{}'x+r_1=0 \tag{12}$$

同理可以类似于定理 1 的如下的定理 2。

定理 2 第 2 类原型超平面的最优方向 w_2 是矩阵 $(S_1-\mu S_2)$ 最大特征值对应的特征向量。

其证明过程完全同于定理 1。同样可以确定第 2 类原型超平面：

$$w_2{}^*{}'x+r_2=0 \tag{13}$$

其中，$r_2=\frac{1}{n_2}\sum_{i=1}^{n_2}w_2{}^*{}'x_i^2$。

从定理 1、定理 2 中可直观地看出，与 GEPSVM 不同，SMGEPSVM 只需要求解一个特征值问题，即用一个 Matlab 命令：$eig(S_1-\mu S_2)$，来获得 2 个原型超平面的最优投影方向。

4 非线性 SMGEPSVM

为了方便地说明非线性 SMGEPSVM，本文定义映射 ϕ 为 $R^d\to H$（核特征空间）。在特征空间的输入样本可以用 $\varphi(X_i)=(\varphi(x_1^i),\varphi(x_2^i),\cdots,\varphi(x_{n_i}^i)),\ i=1,2$ 表示，记 i 是类标。根据线性优化规则，可以考虑下面的非线性 SMGEPSVM 最优准则：

$$\begin{aligned}&\min\ (w_1{}'(S_1-\mu S_2)w_1)^\phi\\&\text{s.t}\ \|w_1^\phi\|=1\end{aligned} \tag{14}$$

特征空间权向量 $w\in H$ 可以由训练样本线性表示[7]，即：

$$w_1=\sum_{i=1}^{n}a_i^1\varphi(x_i)=\varphi(X)a^1 \tag{15}$$

其中，$\varphi(X)=(\varphi(x_1),\varphi(x_2),\boldsymbol{L},\varphi(x_m))$; $a^1=(a_1^1,a_2^1,\cdots,a_m^1)^t$，$m$ 是训练样本数目。则：

$$\begin{aligned}&(w_1{}'(S_1-\mu S_2)w_1)^\phi=(a^1)'\varphi(X)'(\sum_{i=1}^{n_1}x_i^1x_i^1{}'-\mu\sum_{i=1}^{n_2}x_i^2x_i^2{}')\varphi(X)a^1=\\&(a^1)'\varphi(X)'(\varphi(X_1)\varphi(X_1)'-\mu\varphi(X_2)\varphi(X_2)')\varphi(X)a^1=\\&(a^1)'\left(\varphi(X)'\varphi(X_1)\varphi(X_1)'\varphi(X)-\mu\varphi(X)'\varphi(X_2)\varphi(X_2)'\varphi(X)\right)a^1=\\&(a^1)'(K_{m\times n_1}K'_{m\times n_1}-\mu K_{m\times n_2}K'_{m\times n_2})a^1=\\&(a^1)'Ma^1\end{aligned} \tag{16}$$

其中，$M=K_{m\times n_1}K'_{m\times n_1}-\mu K_{m\times n_2}K'_{m\times n_2}$。对于非线性 SMGEPSVM 可以进一步地得到：

$$\begin{aligned}&\min\ \ (a^1)^t Ma^1\\&\text{s.t}\ \ \ \|a^1\|=1\end{aligned} \tag{17}$$

同理，能得到对应优化问题式(6)的非线性模型，如下：

$$\begin{aligned}&\max\ \ (a^2)^t Ma^2\\&\text{s.t}\ \ \ \|a^2\|=1\end{aligned} \tag{18}$$

可以得到核空间的阈值：

$$r_1=\frac{1}{n_1}(a^1)'K_{m\times n_1}1_{n_1},\ \ r_2=\frac{1}{n_2}(a^2)'K_{m\times n_2}1_{n_2}$$

这里，针对非线性 SMGEPSVM 给出另外 2 个类似定理 1 与定理 2 的 2 个定理：

定理 3 核空间中的第 1 类超平面的最优方向 $(w_1)^\phi$ 是矩阵 M 最小特征值对应的特征向量。

定理 4 核空间中的第 2 类超平面的最优方向 $(w_2)^\phi$ 是矩阵 M 最大特征值对应的特征向量。

5 实验结果与分析

对于不同的非线性 SVM 算法，本文用 GAUSS，即 $k(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\gamma^2)$，在 Matlab 7.1, Windows XP 上运行，其内存为 512 MB。

为了说明 SMGEPSVM 分类器的有效性及算法的优越性能，本文在人工数据集与 UCI 数据集上做了实验。表 1 是线性 GEPSVM[6]和 SMGEPSVM 的测试精度比较结果。对于线性问题，所有 2 个算法都只有一个参数：δ 属于 SMGEPSVM 和 GEPSVM。为了得到最好的泛化性能，这个参数在 $\{10^i\mid i=-7,-6,\cdots,+7\}$ 范围中通过调节选择。表 1 中的 Cross planes 数据集是文献[6]中的交叉数据集；“*”指在该数据集上，GEPSVM 中的 G, H 矩阵是奇异的。

表 1　线性 GEPSVM[6]和 SMGEPSVM 测试精度比较

数据集大小：样本数×维数	GEPSVM/(%)	SMGEPSVM/(%)
Cross planes：124×2	100.0	100.0
HILL：606×101	48.2	50.3
CMC：1 473×8	45.9	59.0
SPECT：267×45*	48.1	92.0
LUC：32×57*	72.2	72.2
SMR：208×60*	53.1	56.8
Musk1：476×168*	50.1	62.7

可以看出，在 Cross planes 数据集上，SMGEPSVM 的测试精度与 GEPSVM 相同，而 SMGEPSVM 在真实数据集上的分类性能明显好于 GEPSVM。

表 2 给出了 GAUSS 核 SMGEPSVM, GEPSVM[6]的结果比较。在 2 个算法中，核参数 γ 是在 $\{10^i \mid i=-3,-2,-1,0,+1\}$ 范围中选择。参数 δ 在 $\{10^i \mid i=-7,-6,\cdots,+7\}$ 范围中选择。参数是通过交叉验证获得的。

表 2　非线性 GEPSVM[6]和 SMGEPSVM 测试精度比较

数据集大小：样本数×维数	GEPSVM/(%)	SMGEPSVM/(%)
Cross planes：124×2	100.0	100.0
YEAST：1 484×8	69.1	69.1
MONK1：432×6	63.2	62.5
PIDD：768×8	65.8	68.1
Arrhy：452×279	80.0	80.0

可以看出，SMGEPSVM 有好于 GEPSVM 的性能。与期望相同，非线性 SMGEPSVM 的分类性能也不亚于 GEPSVM。2 个算法的分类时间如表 3 所示。

表 3　线性 GEPSVM[6]和 SMGEPSVM 平均训练时间比较 s

数据集	GEPSVM	SMGEPSVM
Musk1	0.30	0.03

可以看出，SEGEPSVM 的学习时间远远快于 GEPSVM。主要因为：(1)SMGEPSVM 求解的是简单特征值问题；(2)SMGEPSVM 仅需求解 1 个简单特征值问题，而 GEPSVM 需要求解 2 个。

6　结束语

本文提出一个改进型 GEPSVM 算法——SMGEPSVM。该算法通过求解一个标准特征值问题获得 2 个较好的超平面，提高了学习速度，使其在解决 XOR 问题上能获得与 GEPSVM 相当的性能。在真实数据集上，本文方法较 GEPSVM 具有更好的分类性能。但该方法只针对二分类问题提，下一步的研究方向是多类 SMGEPSVM 设计及其相关应用。

参考文献

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York, USA: Springer-Verlag, 1995.

[2] Cherkassky V, Mulier F. Learning from Data: Concept, Theory and Method[M]. New York, USA: John Wiley & Sons, 1997.

[3] Cristianini N, Taylor J S. 支持向量机导论[M]. 李国正, 王　猛, 曾华军, 译. 北京: 电子工业出版社, 2004.

[4] Kojima M, Mizuno S, Noma T, et al. A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems[M]. Berlin, Germany: Springer Verlag, 1991.

[5] Fung G, Mangasarian O. Proximal Support Vector Machine Classifiers[C]//Proc. of KDD'01. San Francisco, CA, USA: [s. n.], 2001.

[6] Mangasarian O, Wild E. MultisurFace Proximal Support Vector Machine Classification via Generalized Eigenvalues[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(1): 69-74.

[7] Mika S, Ratsch G, Weston J. et al. Fisher Discriminant Analysis with Kernels[C]//Proc. of IEEE International Workshop on Neural Networks for Signal Processing. Madison, Wisconsin, USA: [s. n.], 1999.

编辑　顾姣健

（上接第 182 页）

利用 SIFT 旋转不变量算子检测图像并进行匹配时，芯片上字符标志的特点导致匹配特征点时会找到很多不在模板上的点，但其旋转不变特征与模板上的一些特征点重合，因此，将造成无法准确判断模板是否存在于待检测图像中。可见，与 SIFT 相比，VSCPT 的整体判断能力较强。由于模板上的潜在匹配点很多，需要和待测图像对比才能得出符合条件的匹配点，因此不能提出一种标准的模板阈值标准，这是 SIFT 不适用于芯片标记检测的原因之一。

在处理时间上，针对不同偏转角度，利用 VSCPT 和 SIFT 对同一个兴趣区域进行处理时，VSCPT 花费的时间更少、效率更高(表 1)。

表 1　不同偏转角度下 2 种算法的处理时间　ms

算法	-90°	-45°	0°	30°	45°	90°	180°
SIFT	2 500	2 516	2 141	2 328	2 172	2 640	2 297
VSCPT	812	1 343	453	1 703	485	1 031	1 328

6　结束语

本文方案具有一定抗旋转性、抗干扰性和准确性，但可以仍然能加以改进，具体如下：

(1)先在图像的低分辨率层面进行粗匹配，再对候选点利用泽尔尼克算子进行精确匹配，以进一步提高准确性。

(2)当半径较小时，统计环带内的像素灰度值，以提高抗旋转能力。

(3)通过以下操作提高处理速度：1)对图像进行二值化，统计圆内白点所占的比例，粗略求出备选中心点，并利用 VSCPT 法进行检测；2)将实模板和虚模板的匹配过程转换到频域上，把时域卷积换成频域乘积处理。

(4)对一维特征向量进行归一化处理，以提高对亮度的抗干扰能力。

参考文献

[1] 梁志贞, 施鹏飞, 周煦潼. 工业器件上的字符提取及识别[J]. 计算机工程, 2005, 31(9): 41-42.

[2] Prokop R J, Reeves A P. A Survey of Moment-based Techniques for Unoccluded Object Representation and Recognition[J]. Graphical Models Image Process, 1992, 54(5): 438-460.

[3] Lin Yi-Heien, Chen Chin-Hsing. Template Matching Using the Parametric Template Vector with Translation, Rotation and Scale Invariance[J]. Pattern Recognition, 2008, 41(7): 2413-2421.

[4] Choi Min-Seok, Kim Whoi-Yul. A Novel Two Stage Template Matching Method for Rotation and Illumination Invariance[J]. Pattern Recognition, 2002, 35(1): 119-129.

编辑　陈　晖