

最值间距支持向量机

王至超 张化祥

(山东师范大学信息科学与工程学院 济南 250014)

(山东省分布式计算机软件新技术重点实验室 济南 250014)

摘 要 GEPSVM(Proximal Support Vector Machine Classification via Generalized Eigenvalues)是近年提出的一种新的二分类 SVM,其核心思想是通过求解广义特征方程得到两个最优超平面,然后通过计算样本到超平面的距离来决定样本所属类别。与传统 SVM 相比,GEPSVM 降低了时间复杂度,但仍存在奇异性等问题。提出了一种新的算法 TDMSVM(Twin Distance of Minimum and Maximum Support Vector Machine),其通过求解标准特征方程得到两个最优超平面,使超平面满足到本类样例的平均距离最小化,同时到另一类样例的平均距离最大化。通过理论分析和实验证明,与 GEPSVM 相比,TDMSVM 有以下优势:进一步降低了时间复杂度;不需引入正则项,从而提高了泛化性能;克服了奇异性。

关键词 模式识别,特征向量,支持向量机,拉格朗日乘子法

中图分类号 TP18 **文献标识码** A

Twin Distance of Minimum and Maximum Support Vector Machine

WANG Zhi-chao ZHANG Hua-xiang

(Department of Information Science and Engineering, Shandong Normal University, Jinan 250014, China)

(Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan 250014, China)

Abstract GEPSVM is a newly proposed binary SVM in recent years. It learns two optimal hyperplanes by solving the generalized eigenequation and determines the categories of patterns based on the distances between test sample and the two hyperplanes. Compared with traditional SVM, GEPSVM can reduce the time complexity, but it still leaves singularity issues unsolved. This paper introduced a new algorithm TDMSVM(Twin Distance of Minimum and Maximum Support Vector Machine) which seeks two optimal hyperplanes through solving the standard eigenequation and requires the minimal average distance between the same class of samples and hyperplanes. Compared with GEPSVM, TDMSVM has the following advantages: it further reduces the time complexity and does not require the introduction of regular items which improves the generalization performance and avoids singularity.

Keywords Pattern recognition, Eigenvector, Support vector machine, Lagrange multiplier method

1 引言

Vapnik 等人在统计学理论中的结构风险最小化和 VC 维的基础上提出支持向量机(SVM)^[1]。传统的 SVM 方法在不等式约束条件下通过求解二次规划得到最优超平面,在数据线性可分的条件下将两类数据正确分开,并且使分类间隔最大,然而样本数目较大时训练 SVM 比较费时^[2,3]。

Fung 和 Mangasarian 提出的 PSVM^[4]将传统 SVM 中的不等式约束替换为等式约束,通过求解线性方程组得到两个平行的超平面。与传统 SVM 相比,PSVM 的时间复杂度得到了改进,但是在处理 XOR 问题时,仍然没有很好地解决数据“交叉”带来的问题。为了克服此缺点, Mangasarian 和 Wild 在此基础上摒弃了平行的约束条件,提出 GEPSVM^[5]。

其核心思想是,每个超平面离本类样本尽可能近,离另一类样本尽可能远,通过求解广义特征方程可以得到两个最优超平面。GEPSVM 的优势是不需要求解二次规划问题,所需时间少于传统 SVM 并且能很好解决 XOR 问题。近年来,基于 GEPSVM 又发展出多种 SVM 算法^[6-9]。文献[6,7]仍是基于求解广义特征方程,文献[8,9]涉及到矩阵求逆操作,所以它们仍然存在奇异性。杨和陈等人将 GEPSVM 的概念推广到局部化 GEPSVM^[10]和半监督 SemiGEPSVM^[11]。局部化 GEPSVM 根据超平面上的局部凸区域决定样本的类别,但是仍避免不了正则化因子的选择问题。

通过以上分析,GEPSVM 仍然存在如下不足:(1)求解广义特征方程时所涉及的矩阵可能是半正定矩阵,从而导致奇异性;(2)求解广义特征方程得到最优解,时间复杂度 $O(n^3)$

到稿日期:2011-05-03 返修日期:2011-08-01 本文受国家自然科学基金项目(61170145),山东省自然科学基金项目(ZR2010FM021),山东省科技攻关计划项目(2010G0020115),山东省电子发展基金项目(2008B0026),山东省教育厅科研项目(J09LG02)资助。

王至超(1987-),男,硕士生,主要研究方向为数据挖掘、机器学习,E-mail:lsws33@yahoo.com.cn;张化祥(1966-),男,教授,博士生导师,主要研究方向为机器学习、模式识别及 Web 挖掘等,E-mail:huaxzhang@163.com(通信作者)。

仍偏高; (3) 在真实数据集中分类正确率稍低于 SVM^[8]。为解决上述问题, 本文提出基于标准特征方程的最值间距 TDMSVM。与 GEPSVM 相比, TDMSVM 有以下优点: (1) 通过求标准特征方程得到超平面, 因此不存在奇异性问题; (2) TDMSVM 时间复杂度 $O(n^2)$ 小于 GEPSVM; (3) 不需引入 Tikhonov 正则项, 故保留了原始平面的几何意义。

2 GEPSVM 简介

在 n 维空间 R^n 上, 给定两类样本 $\{x_i\}, i=1, 2$, 且分别由矩阵 A 和 B 表示, 样本个数分别为 N_1 和 N_2 。GEPSVM 的目标是在 n 维空间中寻找两个超平面

$$Aw_1 + eb_1 = 0, Bw_2 + eb_2 = 0 \quad (1)$$

使两个超平面满足最小化到本类样本的距离, 最大化到另一类样本的距离。A 类超平面优化准则如下

$$\min_{(w_1, b_1) \neq 0} \frac{\|Aw_1 + eb_1\|^2 / \|[w_1, b_1]^T\|^2}{\|Bw_1 + eb_1\|^2 / \|[w_1, b_1]^T\|^2} \quad (2)$$

式中, A 和 B 分别是大小为 $N_1 \times n$ 和 $N_2 \times n$ 的矩阵, $\|\cdot\|$ 是 L_2 范数, e 是全 1 的列向量。为了避免奇异性, 引入 Tikhonov 正则项, 得到如下形式

$$\min_{(w_1, b_1) \neq 0} \frac{\|Aw_1 + eb_1\|^2 + \delta \|[w_1, b_1]^T\|^2}{\|Bw_1 + eb_1\|^2} \quad (3)$$

$\delta > 0$ 是正则化因子。式(3)被重新定义为如下瑞利熵:

$$\min_{z \neq 0} \frac{z^T L z}{z^T M z} \quad (4)$$

式中, $L, M \in R^{(n+1) \times (n+1)}$ 为对称矩阵:

$$L = [A \ e]^T \times [A \ e] + \delta \times I \quad (5)$$

$$M = [B \ e]^T \times [B \ e], z = [w_1 \ b_1]^T$$

通过求解广义特征方程得到问题(4)的解, 形式如下

$$Lz = \mu Mz, z \neq 0 \quad (6)$$

求解广义特征值问题时 L 和 M 不能同时为半正定矩阵, 否则会产生奇异性。因此, 正则化后的 L 为正定矩阵, 但由此得到的超平面在某种程度上失去了原有的几何意义。准则(2)的全局最优解是式(6)的最小特征值 $\mu_{\min}$ 对应的特征向量, 因此如果 z_1 是最小特征值 $\mu_{\min}$ 对应的特征向量, 那么 $[w_1, b_1]^T = z_1$ 决定了第一类超平面 $x^T w_1 + b_1 = 0$ 。同理, $[w_2, b_2]^T = z_2$ 决定了第二类超平面 $x^T w_2 + b_2 = 0$ 。

3 TDMSVM

与 GEPSVM 的几何意义类似, TDMSVM 的目的也是在 n 维空间中寻找两个如式(1)所示的超平面, 但 TDMSVM 要求出每个超平面到本类样本距离之和的最小均值, 以及到另一类样本的距离之和的最大均值。TDMSVM 与 GEPSVM 有本质的不同, GEPSVM 通过求矩阵的广义特征方程来确定最优超平面。通过下面的分析可知, TDMSVM 通过求矩阵的标准特征方程来确定最优超平面。TDMSVM 使用不同于 GEPSVM 的优化函数, 优化准则如下

TDMSVM1:

$$\min_{w_1, b_1} \frac{1}{2N_1} (Aw_1 + e_1 b_1)^T (Aw_1 + e_1 b_1) -$$

$$\frac{\eta}{2N_2} (Bw_1 + e_2 b_1)^T (Bw_1 + e_2 b_1)$$

s. t. $w_1^T w_1 = 1$

TDMSVM2:

$$\min_{w_2, b_2} \frac{1}{2N_2} (Bw_2 + e_2 b_2)^T (Bw_2 + e_2 b_2) -$$

$$\frac{\eta}{2N_1} (Aw_2 + e_1 b_2)^T (Aw_2 + e_1 b_2)$$

s. t. $w_2^T w_2 = 1$

(8)

以 TDMSVM1 为例, 目标函数第一项是第一类样本到第一类超平面的平均距离, 第二项是第二类样本到第一类超平面的平均距离, A 和 B 分别是大小为 $N_1 \times n$ 和 $N_2 \times n$ 的矩阵, 参数 $\eta \geq 0, e_i = [1, 1, \dots, 1]^T_{m_i \times n}, i=1, 2$ 。依据上述目标函数定义, 得到如下结论。

定理 1 求 TDMSVM 优化问题的最优解 $[w, b]^T$ 可转化为求矩阵 P 的最小特征值对应的特征向量, 其中 P 由同类样本和异类样本协方差矩阵确定。

证明: 同样以 TDMSVM1 为例, 构造 Lagrange 函数:

$$L(w_1, b_1, \lambda) = \frac{1}{2N_1} (Aw_1 + e_1 b_1)^T (Aw_1 + e_1 b_1) - \frac{\eta}{2N_2} (Bw_1 + e_2 b_1)^T (Bw_1 + e_2 b_1) - \lambda(w_1^T w_1 - 1) \quad (9)$$

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

$$\begin{aligned} & \frac{1}{N_1} A^T (Aw_1 + e_1 b_1) - \frac{\eta}{N_2} B^T (Bw_1 + e_2 b_1) - 2\lambda w_1 \\ &= \left(\frac{1}{N_1} A^T A - \frac{\eta}{N_2} B^T B \right) w_1 + \left(\frac{1}{N_1} A^T e_1 - \frac{\eta}{N_2} B^T e_2 \right) b_1 - 2\lambda w_1 \\ &= 0 \end{aligned} \quad (10)$$

令 $\frac{\partial L}{\partial b_1} = 0$, 整理得

$$\begin{aligned} & \frac{1}{N_1} e_1^T (Aw_1 + e_1 b_1) - \frac{\eta}{N_2} e_2^T (Bw_1 + e_2 b_1) \\ &= \left(\frac{1}{N_1} e_1^T A + \frac{\eta}{N_2} e_2^T B \right) w_1 + \left(\frac{1}{N_1} e_1^T e_1 - \frac{\eta}{N_2} e_2^T e_2 \right) b_1 \\ &= 0 \end{aligned} \quad (11)$$

由式(11)得

$$b_1 = \frac{N_2 e_1^T A - N_1 \eta e_2^T B}{N_1 \eta e_2^T e_2 - N_2 e_1^T e_1} w_1 \quad (\eta \neq 1) \quad (12)$$

将式(12)代入式(10), 整理得

$$Pw_1 = \pi w_1$$

$$\begin{cases} P = (N_2 A^T A - N_1 \eta B^T B) + \\ \frac{(N_2 A^T e_1 - N_1 \eta B^T e_2)(N_2 A^T e_1 - N_1 \eta B^T e_2)^T}{N_1 \eta e_2^T e_2 - N_2 e_1^T e_1} \\ (\eta \neq 1) \\ \pi = 2\lambda N_1 N_2 \end{cases} \quad (13)$$

易看出, 矩阵 P 的第一项为同类样本协方差矩阵的线性组合, 第二项的分母可视为常数。将分子展开, 得到同类样本和异类样本的协方差矩阵的线性组合。综上所述, 矩阵 P 的简化形式为

$$P = K_1 A^T A + K_2 B^T B + K_3 A^T B \quad (14)$$

式中, $K_1, K_2, K_3 \in R^{n \times n}$ 。

再将式(12)代入式(7), 整理得

$$\min w_1^T P w_1 = \min w_1^T \pi w_1 = \min \pi \quad (15)$$

由式(13)、式(15)可知, 式(7)的目标函数取最小特征值 π 时, 超平面的最优方向为矩阵 P 的最小特征值所对应的特

征向量,其中 P 由同类样本和异类样本协方差矩阵确定。

同理可证, TDMSVM2 具有同样结论。

与 GEPSVM 不同, TDMSVM 只需求解矩阵 P 的标准特征值,不存在奇异性问题,因此不需要引入正则项,从而保留了分类器的几何特性,这对提高分类器的泛化能力有很重要的作用。

为了直观描述 TDMSVM,并与传统 SVM 做对比,可视化的显示方式如图 1 所示。可看到 SVM 目标是寻找一个最优分类超平面,不但能将两类数据正确分开,并且使分类间隔最大化;TDMSVM 的目标是构造两个超平面,以保证每个超平面到本类样本距离之和的均值最小,同时到另一类样本的距离之和的均值最大。

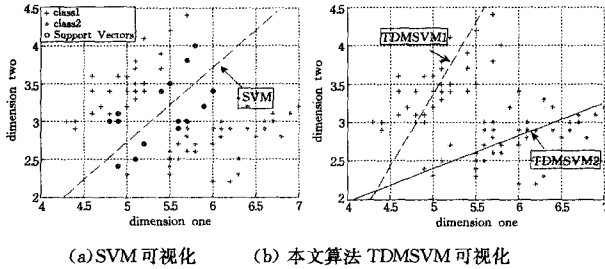


图 1 SVM 和 TDMSVM 可视化

时间复杂度分析:SVM 求二次规划问题时,时间复杂度为 $O(n^{3.5})$ [10];GEPSVM 求广义特征值问题时,时间复杂度为 $O(n^3)$;TDMSVM 求标准特征值问题时,时间复杂度为 $O(n^2)$ 。GEPSVM 和 TDMSVM 的目标是得到两个最优超平面,所以需计算 2 次特征值。假定样本数为 n ,3 种算法的时间复杂度分析如下:

$$\frac{O(\text{SVM})}{O(\text{TDMSVM})} = \frac{n^{3.5}}{2 \times (n/2)^2} = 2n^{1.5} \quad (16)$$

$$\frac{O(\text{GEPSVM})}{O(\text{TDMSVM})} = \frac{2 \times (n/2)^3}{2 \times (n/2)^2} = \frac{n}{2}$$

通过式(16)的分析可看出,与其他两种算法相比,TDMSVM 在时间复杂度方面得到了很大的改善。

4 非线性 TDMSVM

为将 TDMSVM 扩展到非线性,考虑基于核的分类器:

$$K(x^T, C^T)u_1 + b_1 = 0, K(x^T, C^T)u_2 + b_2 = 0 \quad (17)$$

$$C^T = [A \ B]^T$$

K 为任意核函数。与第 2 节类似,构造优化准则如下:

$$\text{KTDSVM1:}$$

$$\text{Min: } \frac{1}{2N_1} (K(A, C^T)u_1 + e_1 b_1)^T (K(A, C^T)u_1 + e_1 b_1) - \frac{\eta}{2N_2} (K(B, C^T)u_1 + e_2 b_1)^T (K(B, C^T)u_1 + e_2 b_1)$$

$$\text{s. t. } u_1^T u_1 = 1 \quad (18)$$

$$\text{KTDSVM2:}$$

$$\text{Min: } \frac{1}{2N_2} (K(B, C^T)u_2 + e_2 b_2)^T (K(B, C^T)u_2 + e_2 b_2) - \frac{\eta}{2N_1} (K(A, C^T)u_2 + e_1 b_2)^T (K(A, C^T)u_2 + e_1 b_2)$$

$$\text{s. t. } u_2^T u_2 = 1 \quad (19)$$

可得如下结论。

定理 2 求 KTDSVM 优化问题的最优解可转化为求

矩阵 S 的最小特征值对应的特征向量,从而得到最优解 $[u, b]^T$ 。

证明:以 KTDSVM1 为例,证明过程与定理 1 类似,得

$$b_1 = \frac{N_2 e_1^T K(A, C^T) - N_1 \eta e_2^T K(B, C^T)}{N_1 \eta e_2^T e_2 - N_2 e_1^T e_1} u_1 \quad (20)$$

$$(\eta \neq 1)$$

$$S u_1 = \pi u_1$$

其中,

$$\begin{cases} S = (N_2 K(A, C^T)^T K(A, C^T) - N_1 \eta K(B, C^T)^T K(B, C^T)) + \\ \frac{\| (N_2 K(A, C^T)^T e_1 - (N_1 \eta K(B, C^T)^T e_2) \|^2}{N_1 \eta e_2^T e_2 - N_2 e_1^T e_1} (\eta \neq 1) \\ \pi = 2\lambda N_1 N_2 \end{cases} \quad (21)$$

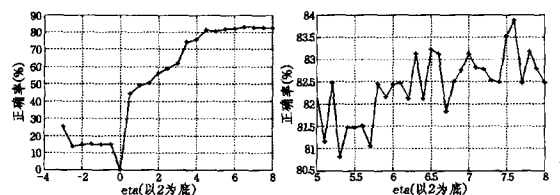
同理, $\min u_1 S u_1 = \min u_1 \pi u_1 = \min \pi$, 即最优解为矩阵 S 的最小特征值对应的特征向量。

5 实验分析

3 种算法的实现基于以下实验平台: Matlab2009a、P4 处理器(2.27G)和 4G RAM,所用数据集来自 UCI 数据集[12]和人工数据集。3 种算法均用十折交叉验证。表 1 是基于线性核 TDMSVM、GEPSVM 和 GunnSVM[13]的结果比较,其中 GunnSVM 是由 Matlab 语言实现的 SVM 工具箱。3 种算法需选择的参数分别为 η 、 δ 和 C 。利用网格搜索算法:最优参数 η 、 δ 和 C 在 $L: \{2^i | i = -3, \dots, 8\}$ 范围中选择,由式(12)可知 $\eta \neq 2^0$ 。如果参数步长设置太小,网格搜索算法会很耗时;步长太大,则容易跳过最优参数。这里采用文献[14]较折中的方法,即先设置较大的步长,粗略估计出次优参数,然后在次优参数的邻近区域搜索最优参数。例如,如图 2 所示 TDMSVM 在线性核条件下寻找 Cleveland Heart 最佳参数 η ,先设置搜索范围 L 的步长为 0.5,搜索到次优的参数 $\eta = 7$ 时,准确率为 82.89%,然后在 2^i 邻近的范围 $L': \{2^i | i = 5, 5.1, \dots, 7.9, 8\}$ 中搜索,缩小步长为 0.1,搜索到最佳的参数 $\eta = 2^{7.6}$ 时,准确率为 83.91%。

表 1 在线性核上的准确率和方差(%)

数据集(大小)	TDMSVM ACC±STD	GEPSVM ACC±STD	GunnSVM ACC±STD
Ionosphere(351×34)	85.55±7.67	78.85±6.09	82.34±6.08
Heart-statlog(270×13)	84.78±6.37	85.21±5.87	84.36±7.1
Votes(435×16)	95.12±3.00	92.66±7.52	94.13±3.35
Cleveland Heart(297×13)	83.91±7.56	81.20±9.65	84.92±6.79
Hepatitis(155×19)	79.29±7.99	80.86±9.77	81.46±7.42
Wpbc(110×32)	68.18±16.80	63.64±9.96	65.84±11.50
Cmc(1473×9)	68.72±5.86	66.23±8.61	67.95±7.75
Breast-w(683×9)	96.61±2.17	88.31±10.13	96.19±1.85
Mean accuracy	82.77	79.62	82.15



(a) 对 η 的最优值进行粗略估计 (步长=0.5, $i \neq 0$) (b) 对 η 最优值进行精细估计 (步长=0.1)

图 2 对 Cleveland Heart 的参数 η 进行估计

表 2 是基于 GAUSS 核 ($\exp(-\sigma * |u-v|^2)$) 的 TDMSVM、GEPSVM 和 GunnSVM 的测试结果比较, 参数 η , δ 和 C 的选择范围与表 1 一致。基于 GAUSS 核的分类器除了要选择以上参数外, 还要选择核参数, σ 在 $\{2^i | i=-5, \dots, 7\}$ 范围中选择, 同样采用网格搜索算法搜索最佳的参数。

表 2 在 GAUSS 核函数上的准确率和方差 (%)

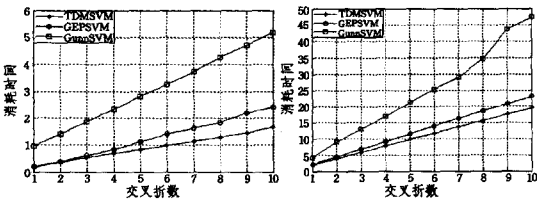
数据集(大小)	TDMSVM ACC±STD	GEPSVM ACC±STD	GunnSVM ACC±STD
Heart-statlog(270×13)	83.16±6.24	84.19±4.69	78.25±7.36
Cleveland Heart(297×13)	84.73±5.63	74.52±9.35	82.32±6.54
Votes(435×16)	95.66±7.72	93.39±8.67	94.21±7.70
Hepatitis(155×19)	81.57±9.06	79.32±9.69	82.38±9.11
Ionosphere(351×34)	91.34±7.79	82.29±10.60	92.16±5.92
Cmc(1473×9)	70.32±5.46	66.73±6.65	71.48±5.60
Wpbc(110×32)	73.22±14.55	62.36±8.11	74.24±6.31
Breast-w(683×9)	95.83±6.89	88.72±5.13	96.74±2.03
Mean accuracy	84.47	78.94	83.97

表 1 及表 2 的结果表明, 在 8 个 UCI 数据集上, TDMSVM 平均模式识别率明显优于 GEPSVM。尽管 TDMSVM 与 GunnSVM 平均模式识别率相当, 但第 2 节式 (16) 的时间复杂度分析和表 3 的实验结果表明, TDMSVM 的模式识别速度更快。

表 3 在线性核上十折交叉验证每一折消耗的平均时间 (s)

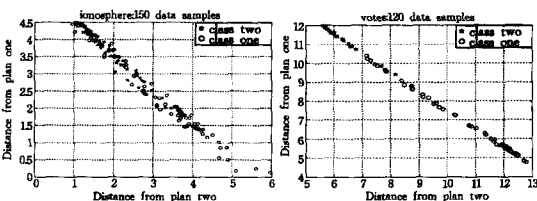
数据集(大小)	TDMSVM	GEPSVM	GunnSVM
Cleveland(297×13)	1.38	1.66	2.76
hapatitis(155×19)	0.40	0.53	0.81
ionosphere(351×34)	1.95	2.32	4.74
Wpbc(110×32)	0.17	0.24	0.52
Heart-statlog(270×13)	1.13	1.26	2.42
Breast-w(683×9)	6.45	9.67	67.90
Cmc(1473×9)	53.61	71.94	386.71
Votes(435×16)	2.69	3.35	24.58

以 wpbc 和 ionosphere 为例, 图 3 是 3 种算法在十折交叉验证中所用的时间对比, x 轴代表交叉折数, y 轴代表随着折数的递增, 时间的累加和。例如, 算法运行到第四折时消耗的时间是前 4 折所用时间的总和。消耗时间随着交叉折数递增而增加, 图 3 也从侧面反映出随着样本个数增多, TDMSVM 在时间上的优势越明显。



(a) 在 wpbc 上所用时间 (b) 在 ionosphere 上所用时间

图 3 3 种算法十折交叉验证所用时间



(a) ionosphere 测试样本到平面的距离 (b) votes 测试样本到平面的距离

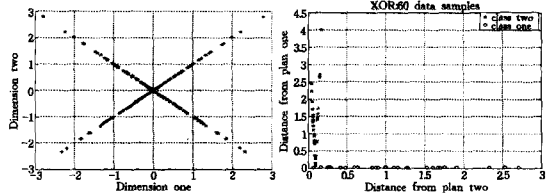
图 4 样本到超平面的距离

在数据集 ionosphere 中随机取 150 个测试样例, 其余数

据作为训练样例。同样从 votes 中随机取 120 个测试样例, 其余数据作为训练样例。图 4 为测试样例到两个超平面的距离, 第一类样例用“○”表示, 第二类样例用“☆”表示, 坐标值表示为 (c_1, c_2) , 其中 $c_1 = |x^T w_1 + b_1|$, $c_2 = |x^T w_2 + b_2|$ 。根据式 (22) 判断测试样例归属于哪一类, 即测试样例被划分到离其距离更近的那一超平面。

$$x^T w^{(r)} + b^{(r)} = \min_{i=1,2} |x^T w^{(i)} + b^{(i)}| \quad (22)$$

与 GunnSVM 比较, GEPSVM 除了在时间上占优势外, 另一优势表现在对 XOR 数据的分类性能上。由于 TDMSVM 的几何意义与 GEPSVM 相似, 因此 TDMSVM 也同样具备这一优势。人工生成 XOR (200×2) 数据集, 数据范围为 $[-3, 3]$, 数据分布关于原点对称, 如图 5(a) 所示。随机取 60 个数据作为测试集, 其余数据作为训练集, 图 5(b) 为基于 GAUSS 核函数, 用于测试数据到 TDMSVM 两个超平面的距离, 第一类样例用“○”表示, 第二类样例用“☆”表示。



(a) XOR 数据分布可视化 (b) XOR 测试样本到平面的距离

图 5 数据可视化和样本到平面的距离

设置 $\eta=0$, 即要求 TDMSVM 的超平面只到本类样本最近。GunnSVM 和 GEPSVM 中要求 C 和 δ 大于 0, 设置 $\delta=C=0.1$ 。3 种算法的核参数选 $1/k$, $k=140$ 为训练样本个数。3 种算法在 XOR 上的准确率如表 4 所列。

表 4 3 种算法准确率对比 (%)

核函数	TDMSVM ACC	GEPSVM ACC	GunnSVM ACC
GAUSS 核	98.33	96.67	76.67
线性核	100.00	95.00	58.33

实验分析: 由表 4 可以看出, 在 XOR 数据集上, TDMSVM 的分类能力同样优于 GEPSVM。GEPSVM 和 TDMSVM 在 XOR 数据集上识别率高的原因如图 5(a) 所示, 这两种算法很容易找到两个超平面来“拟合”XOR 数据集。同样, 由图 5(b) 可看出, 在 XOR 数据集上两类样本到本类超平面的距离几乎为 0, 距离坐标分布几乎不交叉, 说明超平面“拟合”效果很好。经初步分析, 在线性核条件下, GunnSVM 性能远不及其他两种算法的原因是线性核函数在原始空间中特征直接做内积, 并在原始特征空间中寻找一个最优分类超平面, 它不但将两类数据正确分开, 而且使分类间隔最大。由 XOR 数据分布图 5(a) 可看出, 原始空间中很难找到这样的超平面。

结束语 本文提出基于标准特征值问题的 SVM, 它通过求标准解特征值得到最优超平面。经理论证明和实验仿真表明, 它是一种有效的分类算法。与 GEPSVM 相比, TDMSVM 不需要引入正则项, 从而保留了原始平面的几何意义, 并且避免了奇异性问题。本算法在模式识别率和时间上更占优势。下一步工作将从统计角度分析 TDMSVM 的特性。另外, 从本文的实验来看, 线性核与非线性核在样例识别率上互有胜负, 具体选择哪种核函数和相应的参数更适合本文算法, 也将是下一步的研究内容。

参考文献

- [1] Vapnik V N. The Nature of Statistical Learning Theory (seconded) [M]. New York: Springer, 2000
- [2] 文益民, 王耀南, 吕宝粮, 等. 支持向量机处理大规模问题算法综述[J]. 计算机科学, 2009, 36(7): 20-25
- [3] 顾亚祥, 丁世飞. 支持向量机研究进展[J]. 计算机科学, 2011, 38(2): 14-17
- [4] Fung G, Mangasarian O L. Proximal Support Vector Machine Classifiers[C]// Proceedings of Knowledge Discovery and Data Mining. New York, USA: ACM, 2001: 77-86
- [5] Mangasarian O L, Wild E W. Multisurface Proximal Support Vector Classification via Generalized Eigenvalues [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2006, 28(1): 69-74
- [6] Jayadeva, Khemchandani R, Chandra S. Fast and Robust Learning Through Fuzzy Linear Proximal Support Vector Machines [J]. Neurocomputing, 2004, 61: 401-411
- [7] Jayadeva, Khemchandani R, Chandra S. Fuzzy multicategory proximal support vector classification via generalized eigenvalues[J]. Soft Computation, 2007, 17(5): 679-685

- [8] Jayadeva, Khemchandani R, Chandra S. Twin Support Vector Machines for Pattern Classification[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905-910
- [9] Kumar M A, Gopal M. Least squares twin support vector machines for pattern classification[J]. Expert Systems with Applications, 2009, 40(1): 7535-7543
- [10] 杨绪兵, 陈松灿, 杨益民. 局部化的广义特征值最接近支持向量机[J]. 计算机学报, 2007, 30(8): 1227-1234
- [11] 杨绪兵, 潘志松, 陈松灿. 半监督型广义特征值最接近支持向量机[J]. 模式识别与人工智能, 2009, 22(3): 349-353
- [12] Blake C L, Merz C J. UCI Repository for Machine Learning Databases[EB/OL]. <http://www.ics.uci.edu/mllearn/MLRepository.html>, May 2010
- [13] Gunn S R. MATLAB Support Vector Machine Toolbox[EB/OL]. <http://www.isis.ecs.soton.ac.uk/isystems/kernel/>, May 2010
- [14] Hsu C-W, Chang C-C, Lin C-J. A Practical Guide to Support Vector Classification [EB/OL]. <http://www.csie.ntu.edu.tw/~cjlin>, July 2010

(上接第 192 页)

将上述报告用 IBNF 描述如下:

* 状态报告 * * 我方 * * 位置状态 * * 于 * * 20XX 年 3 月 17 日 12 点 24 分 08 秒 * * 到达 * * XX°48'.13N, XX°38'.42E * * 事实情况 * / 报告_引用_标识 1/。

其语法树结构如图 3 所示。

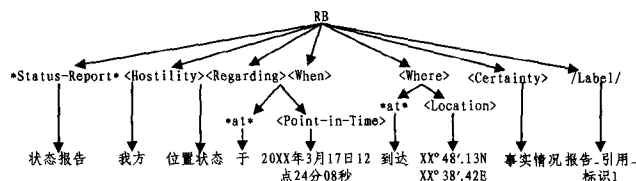


图 3 状态报告语法树

结束语 BML 是一种实现指控与仿真系统间无歧义通信的标准语言规范^[12]。针对典型作战命令的语言结构特点, 在 BNF 基础上进行了适当改进的形式化方法 IBNF, 是用来描述 BML 语法的一种合适选择。

本文提出的 IBNF 语法形式化描述方法能够使作战命令具有结构清晰、规范的统一描述, 使指控系统与仿真系统之间交换的作战命令都具有结构化形式, 从而为实现语法层次上的互操作性提供技术支撑。应用该方法可以构建作战命令模板, 按照模板配置相应参数, 即可生成指控系统和仿真系统都可理解和处理的结构化作战命令, 从而有效实现语法层次的互操作性。

下一步将围绕 BML 体系结构中协议视图和本体视图展开研究, 通过相应的转换算法, 将结构化作战命令转换成符合网络传输需求的标准网络交换形式(如 RDF/XML 文档), 从而实现指控系统与仿真系统之间的数据交换和语义共享, 使指挥控制过程更加迅速、可靠。

参考文献

- [1] Reus D, Kromp D. BML-enabling of national C2 systems for coupling to Simulation[C]// IEEE Spring Simulation Interopera-

- bility Workshop. Providence, Rhode Island, USA, 2008
- [2] 彭勇, 彭春光, 龚建兴, 等. 作战管理语言研究综述[J]. 系统仿真学报, 2009, 21(17): 5336-5339
- [3] Tolk A, Turnitsa C. Battle Management Language: A Triangle with Five Sides [C]// Fall Simulation Interoperation Workshop, 06S-SIW-016. Orlando, FL, USA, 2006
- [4] Andreas T, Muguira J A. The Levels of Conceptual Interoperability Model (LCIM) [C]// Proceedings IEEE Fall Simulation Interoperability Workshop. Orlando, FL: IEEE CS Press, September 2003
- [5] Ma Liang-jun, Chen Lin, Lu Peng, et al. A new knowledge representation method for Battle Management Language[C]// 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Yantai, Shandong, Aug. 2010: 1761-1766
- [6] Ulrich S, Haarmann, Bastian, et al. A Grammar for Battle Management Language[C]// IEEE/ACM 15th International Symposium on Distributed Simulation and Real Time Applications (DS-RT). Salford, United Kingdom, Sept. 2011: 155-159
- [7] Gao Xiao-lei. The Design and Implementation of Z Language Editor[C]// Algorithms and Architectures for Parallel Processing, 9th International Conference (ICA3PP). Taipei, Taiwan, June 2009
- [8] 叶丽君. 基于 UML 描述的概念模型校验技术研究[D]. 郑州: 解放军信息工程大学, 2008
- [9] 顾妍午, 王遵彤, 吴启迪. 面向对象 Petri 网技术在系统建模中的应用[J]. 同济大学学报: 自然科学版, 2010, 38(3): 437-441
- [10] Ye C, Zhong J, Feng Y. Attribute-based access control policy specification language [J]. Journal of Southeast University, 2008, 24(3): 260-263
- [11] Schade U, Michael R. H. Formalizing Battle Management Language: A Grammar for Specifying Orders[C]// 2006 Spring Simulation Interoperability Workshop, 06S-SIW-068. Huntsville, AL, 2006
- [12] 岳秀清, 张磊, 别晓峰. 外军指挥控制建模研究现状及启示[J]. 军事运筹与系统工程, 2009, 23(1): 70-74