一种基于近邻边界的粒度支持向量机学习策略

张春艳 倪世宏 查 翔
(空军工程大学航空航天工程学院 西安 710038)

摘 要 针对粒度支持向量机进行粒划分后提取代表点时丢失部分重要分类信息从而影响分类准确率的情况,提出了一种基于近邻边界的粒度支持向量机(Neighboring-boundary Granular Support Vector Machine,NGSVM)的学习策略。首先采用kmeans方法进行粒划分,对不同的粒依据不同的规则提取粒内代表点,并按照要求分别将代表点放入精简集或修正集中,再用这两个集合中的数据对支持向量机进行训练;形成分类器后,根据核空间距离提取靠近分类面的近邻边界数据以对分类面进行修正。仿真实验结果表明,通过提取靠近分类面的近邻边界数据进行重新训练,能够修正分类面,进一步提高粒度支持向量机的分类准确率。

关键词 近邻边界,粒度支持向量机,粒度,精简集,修正集
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.050

Learning Strategy Based on Neighboring-boundary Granular Support Vector Machine

ZHANG Chun-yan NI Shi-hong ZHA Xiang
(College of Aeronautics and Astronautics Engineering, Air Force Engineering University, Xi'an 710038, China)

Abstract Granular support vector machine will lead to loss of partial classification information and accuracy degradation while dividing granules and extracting representative points. To solve this problem, a learning strategy based on neighboring-boundary granular support vector machine (NGSVM) was proposed. Samples were divided into granules with kmeans method firstly, different granules were dealt with different rules to extract representative points, and then these representative points were put into fixed set or reduced set as requested, by which support vector machine (SVM) was trained. After completion of classifier, classification plane would be rectified by extracting neighboring-boundary samples according to the kernel distance. The simulation results show that NGSVM gains a higher classification accuracy by extracting neighboring-boundary samples near classification plane and fixing classification plane.

Keywords Neighboring-boundary, Granular support vector machine, Granules, Reduced set, Fixed set

1 引言

SVM在处理大规模数据集以及混淆较为严重的数据集时,会出现训练样本多、速度慢、效率低等问题[1]。对此,许多学者做了很多提高SVM训练效率的研究[2,3]。Tang在2004年将粒度计算(Granular Computing,GrC)[4]与支持向量机结合,提出了粒度支持向量机(Granular Support Vector Machine,GSVM)[5]的概念。该方法以聚类的方式对数据进行粒划分,然后取代表点进行训练,有效地缩减了训练数据的规模,提高了训练效率。文献[6]提出了一种基于粒分布的支持向量机加速训练方法(Distribution Based on GSVM,DGSVM),虽然DGSVM极大地压缩了训练样本,缩短了训练时间,但是基于粒划分取代表点进行训练,会丢失对分类器的形成有重要贡献的边界样本,从而降低粒度支持向量机的分类准确率。

为提高粒度支持向量机的分类准确率,文献[7]提出了核空间下的GSVM(KGSVM)模型,解决了粒划分时因核空间数据分布不一致而导致的重要分类信息丢失的问题,但是在核空间中进行粒划分时,超球粒的半径应符合KGSVM_R规则,如若同类或不同类中超球半径差异较大,会影响学习的效果。文献[8]提出了一种基于关联规则的核粒度支持挖掘,通过适当控制频繁度和关联度两个关键参数,获得了较好的分类准确率。

文献[9]在处理不平衡数据时,提出了一种边界样本选择方法,运用这种方法构建一个随机小样本池,在选择边界样本的过程中不需要在全部训练集范围内进行搜索,只需要在随机选择的部分样本中进行搜索。文献[10]在学习的前半部分使用了加权压缩近邻方法来选择最具代表性的样本点进行训练,在学习的后半部分使用文献[9]所提到的方法,利用最近边界规则在随机小样本池中选择边界样本来提高分类准确率。

为充分利用边界数据对分类器的贡献,本文在DGSVM的基础上,借鉴文献[9,10]的思想,提出一种基于近邻边界的粒度支持向量机学习策略(Neighboring-boundary Granular Support Vector Machine,NGSVM)。先采用kmeans方法对原数据进行粒划分,计算各个粒的混合度以及平均混合度,若某粒的混合度大于平均混合度,则取代表点放入精简集;反

到稿日期:2015-07-19 返修日期:2015-10-04
张春艳(1987—),女,硕士生,主要研究方向为数据处理与挖掘,E-mail:zhangchy125@163.com;倪世宏(1963—),男,教授,博士生导师,主要研究方向为飞行状态监控与地面数据处理;查 翔(1988—),男,博士生,主要研究方向为人工智能、数据处理与挖掘。

之，则取代表点放入修正集。用精简集训练支持向量机得到初始分类器，再用修正集对分类器进行第一次修正。然后提取靠近分类面的近邻边界数据放入精简集，进行分类器的第二次修正，以期在不明显降低粒度支持向量机训练效率的前提下进一步提高分类准确率。

2 NGSVM 算法

2.1 NGSVM 算法基本思路

先采用 kmeans 方法对数据集 $\boldsymbol{X}$ 中所有数据进行初次粒划分，得到粒组 $\boldsymbol{G}=\bigcup_{r=1}^{k}\{\boldsymbol{G}_r\}$（$k$ 为划分的粒数）。若某粒中只有一类数据，则该粒为纯粒，反之则为混合粒。对于纯粒，取该粒中心点放入修正集；对于混合粒，则按照下述步骤进行计算。

若数据维数为 dim，采用欧氏距离作为两个数据之间的距离，设混合粒 $\boldsymbol{G}_r$ 中数据 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 的坐标为：

$$\boldsymbol{x}_i=(x_i^1,x_i^2,x_i^3,\cdots,x_i^{dim})$$

$$\boldsymbol{x}_j=(x_j^1,x_j^2,x_j^3,\cdots,x_j^{dim})$$

两个数据之间的距离为

$$dist(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{\sum_{s=1}^{dim}(x_i^s-x_j^s)^2} \tag{1}$$

根据混合粒 $\boldsymbol{G}_r$ 中正负两类数据的数量定义粒的混合度[6]：

$$mix=1-\frac{|n_1-n_2|}{n_1+n_2} \tag{2}$$

其中，n_1、n_2 分别表示混合粒 $\boldsymbol{G}_r$ 中正负两类数据的个数，$mix\in(0,1]$，mix 值越高，表明正负类数据规模越相当，mix 值越低则表示两类数据规模悬殊越大。所有混合粒的平均混合度定义为

$$meanmix=\frac{1}{k}\sum_{i=1}^{k}mix(i) \tag{3}$$

再次采用 kmeans 方法，分别将所有混合粒中的正负类数据进行粒划分，划分的粒数分别设定为 $mix\times n_1$、$mix\times n_2$，取划分后的每个粒的中心点作为代表点，判断每个混合粒的混合度是否大于平均混合度。若大于，则将代表点放入到精简集中；若小于，则将代表点放入到修正集中。

分类器的迭代修正用精简集数据对分类器进行训练，用修正集数据对得到的分类器进行测试，将修正集中分类错误的数据移到精简集中。用更新过的精简集分类器再次训练，并用修正集进行测试，直至修正集中的数据全部分类正确。

上述采用提取代表点的方式进行训练会丢失一些支持向量，造成一部分数据的错分。为降低错分的概率，可提取靠近分类面处的近邻边界数据放入到精简集中，进行分类器的第二次修正。最优分类面的判别函数用核函数表示[11]：

$$f(\boldsymbol{x})=\mathrm{sgn}\{\sum_{i=1}^{l}\alpha_i^*\,y_iK(\boldsymbol{x}_i,\boldsymbol{x})+b^*\} \tag{4}$$

式中，α_i^*（$i=1,2,\cdots l$）为最优拉格朗日乘子，$\boldsymbol{x}_i$（$i=1,2,\cdots,l$）为支持向量，y_i（$i=1,2,\cdots,l$）为支持向量对应的类标签，$K(\boldsymbol{x}_i,\boldsymbol{x})=\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x})$ 为核函数，φ 为将数据由原空间映射到特征空间中的映射函数，b^* 为最优偏移量。

数据 $\boldsymbol{x}$ 到最优分类面的距离为[11]：

$$d=\frac{|\sum_{i=1}^{l}\alpha_i^*\,y_iK(\boldsymbol{x},\boldsymbol{x}_i)+b|}{\|w\|} \tag{5}$$

将 d 由小到大排序，取前 m 个近邻边界数据放入到精简集中，与修正集一起进行分类器的第二次修正。

2.2 NGSVM 算法流程

根据 NGSVM 算法的基本思路，NGSVM 的算法流程分为两个阶段：第一阶段完成粒的划分和代表点的选取，第二阶段完成支持向量机的训练及修正，如图 1 所示。

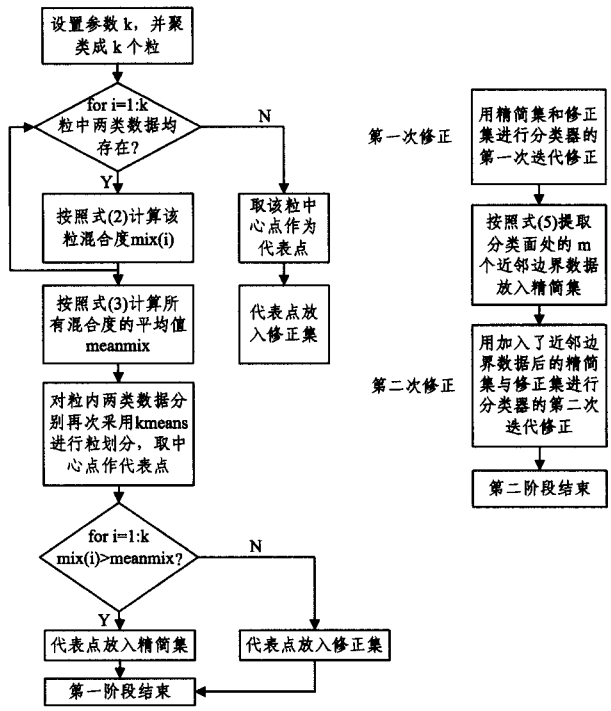


图 1 NGSVM 算法流程

3 实验结果及分析

仿真实验软件为 Matlab R2009a，支持向量机工具箱为 lssvm，训练时采用 RBF 核函数，惩罚因子 C 取值 200，核参数 σ^2 取值 15。实验所选的 4 个数据集（见表 1）中，waveform-40、shuttle、banana 来自 UCI 数据集，adult 来自 lssvm 工具箱中自带的测试数据。因支持向量机是针对二分类问题的，waveform-40 有 3 类数据，取类标签为 1 和 2 的数据进行训练测试；shuttle 有 9 类数据，取类标签为 1 和 4 的部分数据进行训练测试。

表 1 实验数据集

数据集	训练集规模(正/负)	测试集规模(正/负)	数据维数
waveform-40	1800(882/918)	1543(771/773)	40
adult	2418(460/1958)	1158(223/935)	14
shuttle	3000(1429/1571)	1194(627/567)	9
banana	8800(3893/4907)	1000(470/530)	2

3.1 近邻边界数据对分类面的影响

现以 banana 数据集为例，讨论所提取的近邻边界数据对分类面的影响。为避免因数据太多而影响观察与分析，取 banana 训练数据中前 1000 个进行训练，取 banana 数据集的全部测试数据进行测试，粒划分个数 $k=30$，提取的近邻边界数据个数 $m=100$。

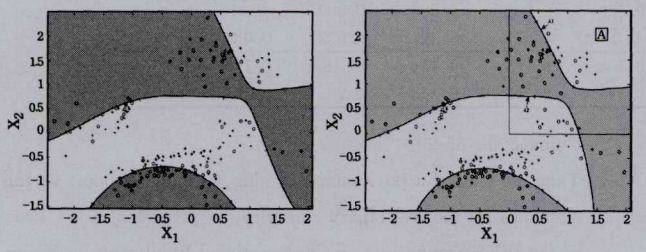
表 2、表 3 列出了 NGSVM 的第一次、第二次修正过程中精简集、修正集和分类准确率的变化情况。图 2、图 3 为 NGSVM 的第一次、第二次修正过程中分类面的变化情况，星形为正类支持向量，圆圈为负类支持向量，黑色线条表示的是最优分类面。图 3 中，近邻边界正类数据用被正方形包裹住的星形表示，近邻边界负类数据用被圆圈包裹住的加号表示。

表 2　NGSVM 第一次修正过程数据

第一次修正	精简集规模	修正集规模	分类准确率/%
初始状态	191	30	76.67
1 次迭代	198	23	95.65
2 次迭代	199	22	100

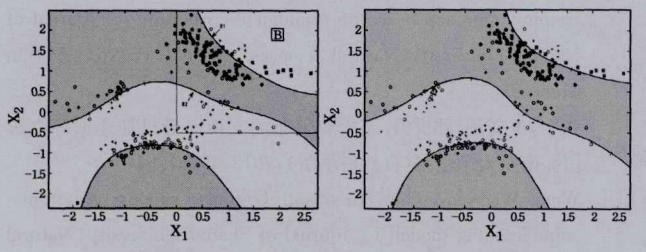
表 3　NGSVM 第二次修正过程数据

第二次修正	精简集规模	修正集规模	分类准确率/%
初始状态	299	22	90.91
1 次迭代	301	20	95
2 次迭代	302	19	100



(a)初始分类面　　(b)2 次迭代后分类面

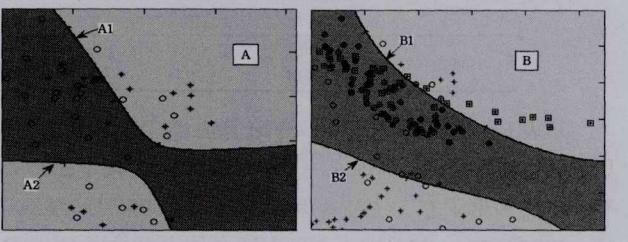
图 2　NGSVM 的第一次修正过程分类面



(a)初始分类面　　(b)2 次迭代后分类面

图 3　NGSVM 的第二次修正过程分类面

从表 2 和表 3 可看到，NGSVM 第一次修正结束与第二次修正开始时，精简集中多出了根据式(5)提取的前 100 个数据。从对应的图 2(b)、图 3(a)可看到，这 100 个近邻边界数据的加入使训练得到的分类面有了明显的变化。图 4(a)为图 2(b)A 区域局部放大图，图 4(b)为图 3(a)B 区域的局部放大图，图 4(b)比图 4(a)多出了用被正方形包裹住的星形和被圆圈包裹住的加号的数据，这表明在 100 个近邻边界数据中，包含有粒划分提取代表点时丢掉的支持向量。近邻边界数据的加入使支持向量机沿着支持向量的分布确定分类面，从而使得分类面 A_1 移至 B_1，分类面 A_2 移至 B_2。通过分析比较可以看出，在先粒划分再取代表点的情况下，提取近邻边界数据加入分类器的训练能够改变分类面，从而改善分类效果。



(a) 图 2(b)A 区域放大图　　(b) 图 3(a)B 区域放大图

图 4　区域放大图

3.2　近邻边界数据个数 m 对分类准确率的影响

表 4 列出了 4 个数据集在粒划分个数 $k=30$ 及取近邻边界数据 $m=10$ 至 100 时，NGSVM 与 DGSVM 分类准确率的对比。由表 4 中数据可看出：(1)当提取近邻边界数据加入训练后，相比于 DGSVM 算法，NGSVM 算法的分类准确率有不同程度的提高；(2)当 m 增大到一定程度时，分类准确率趋于稳定。因每个数据集中数据分布都不尽相同，m 不是越大分类准确率越高。因为支持向量机训练的过程实际就是在边界样本中寻找支持向量从而确定最优分类面的过程，当 m 增大时，如果提取的数据中包含的粒划分时丢失的支持向量没有增加，则这些数据对分类面的改变贡献不大，分类准确率也不会有较大变化。

表 4　近邻边界数据个数对分类准确率的影响($k=30$)

k=30	DGSVM	NGSVM(k=30)									
		m=10	m=20	m=30	m=40	m=50	m=60	m=70	m=80	m=90	m=100
banana	75.9	77.8	84.2	79.6	78.9	77.6	77.4	78.4	77.7	80.3	80.5
waveform-40	88.15	89.9	83.36	87.31	88.48	88.35	90.94	90.48	90.48	90.61	90.68
shuttle	88.03	87.36	92.13	94.56	98.25	97.41	94.31	96.4	97.24	97.24	96.65
adult	76.34	76	77.38	76.95	78.5	79.37	78.33	79.54	79.8	80.57	79.88

3.3　粒划分个数 k 对分类准确率的影响

表 5 列出了 4 个数据集在 k 变化时的分类准确率，选取的近邻边界数据个数 $m=100$。从表 5 中可看到，当粒划分个数 k 增大时，对 banana、waveform-40、shuttle 这 3 个数据集而言，分类准确率逐渐增大，当 k 增大到一定程度时，分类准确率逐渐稳定，并接近 SVM 的分类准确率。对 adult 数据集而言，分类准确率没有显著变化。

当 k 较小时($k\leqslant 30$)，k 个粒的数据规模相对较大，所提取的粒内代表点虽然能够较好地代表该粒中的数据，但因选取的代表点分布比较集中，未能充分代表所有数据的分布特征，导致分类准确率较低。当 k 较大时，k 个粒分布在原空间的各个角落，粒中数据规模相对较小，所提取的粒内代表点更能代表所有数据的分布情况，故分类准确率会增大。但对于 adult 数据集，k 增加后，虽然粒分布较为分散，但所提取的负类数据居多，最终训练数据不均衡，SVM 训练时将少数的正类数据归为负类[12]，导致分类准确率提高不明显。

表 5　k 对分类准确率的影响(单位：%)

		k=10	k=20	k=30	k=40	k=50	k=60	k=70	k=80	SVM
banana	DGSVM	84	85.7	75.9	85.7	88.4	88.8	89.4	89.5	87.7
	NGSVM	84.5	87.6	80.5	88.1	88.7	89.4	90	89.7	
waveform-40	DGSVM	81.68	88.28	88.15	88.61	89.64	91	90.23	89.97	91.13
	NGSVM	89.71	91.59	90.68	91.33	90.87	91.33	88.54	91.84	
shuttle	DGSVM	81.24	93.47	88.03	95.48	95.98	97.32	93.72	97.91	99.58
	NGSVM	96.57	96.49	96.65	95.31	93.97	96.15	95.31	96.24	
adult	DGSVM	78.07	78.93	76.34	79.88	77.64	77.73	78.16	77.81	84.89
	NGSVM	80.92	78.16	79.88	81.01	78.42	80.23	80.23	77.47	

3.4 NGSVM 的时间开销

表 6 列出了 SVM、DGSVM、NGSVM 3 种算法在 4 个数据集上的训练所耗时间。表 6 中的时间包括了粒划分的时间、提取近邻边界数据的时间和支持向量机的训练时间。实验中硬件配置：CPU 为 Intel Core i3-3220，主频为 3.3GHz，内存为 1.97GB。

表 6　3 种方法的时间对比(单位：s)

		k=10	k=20	k=30	k=40	k=50	k=60	k=70	k=80	SVM
banana	DGSVM	0.859	0.422	0.665	0.473	0.739	0.849	0.62	0.616	94.25
	NGSVM	0.975	0.518	0.762	0.557	0.801	0.938	0.706	0.716	
waveform-40	DGSVM	0.113	0.117	0.119	0.124	0.112	0.125	0.115	0.137	1.38
	NGSVM	0.163	0.168	0.171	0.177	0.164	0.178	0.165	0.192	
shuttle	DGSVM	0.082	0.037	0.01	0.008	0.008	0.008	0.009	0.01	3.13
	NGSVM	0.18	0.078	0.032	0.03	0.031	0.03	0.032	0.036	
adult	DGSVM	0.17	0.15	0.138	0.138	0.165	0.154	0.18	0.199	10
	NGSVM	0.272	0.265	0.24	0.242	0.266	0.263	0.286	0.295	

从表 6 可看出，相比于 DGSVM 算法，NGSVM 增加了提取近邻边界数据进行训练这一步骤，故在训练时间上比 DGSVM 算法多出了 0.1s 左右，但比 SVM 的训练时间缩减了 10～100 倍左右。这是由于 NGSVM 总体上继承了 DGSVM 训练速度快的优点。在某些工程领域，存在对分类准确率和训练速度均有一定的要求且对前者的要求要高于后者的情况，此时，在不显著降低训练速度的前提下研究分类准确率更高的算法，在计算机运行速度日新月异的今天仍有较大的现实意义。

结束语　为提高粒度支持向量机的分类准确率，本文提出一种基于近邻边界的粒度支持向量机(NGSVM)学习策略，该算法的核心是将提取的近邻边界数据加入精简集以对分类器进行训练。仿真实验结果表明，通过提取近邻边界数据可适度提高粒度支持向量机的分类准确率；同时本算法继承了 DGSVM 训练速度快的优点，训练所耗时间增加不多，具有一定的工程实用价值。本文仅对二分类问题利用中等规模的数据集这种情况进行了仿真验证，在以后的工作中，应将本文所提算法扩展到多类分类问题中，并在大规模数据集上进行深入的仿真验证。

参 考 文 献

[1] Ding Shi-fei, Qi Bing-juan, Tan Hong-yan. An overview on theory and algorithm of support vector machines[J]. Journal of University of Electronic Science and Technology of China, 2011, 40(1): 2-10(in Chinese)
丁世飞，齐丙娟，谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报，2011，40(1)：2-10

[2] de Almeida M B, de Padua Braga A, Braga J P. SVM-KM: speed SVMs learning with a priori cluster selection and k-means[C]// Proceedings of the Sixth Brazilian Symposium on Neural Networks. Riode Janeiro, R J, Brazil, 2000: 162-167

[3] Horng S-J, Su Ming-yang, Chen Y-H, et al. A novel intrusion detection system based on hierarchical clustering and support vector machines [J]. Expert Systems with Applications, 2011, 38(1): 306-313

[4] Yao Yi-yu. Perspectives of granular computing[C]// Proc of the IEEE International Conference on Granular Computing. Beijing, China, 2005: 85-90

[5] Tang Yu-chun, Jin Bo, Zhang Yan-qing. Granular support vector machines for medical binary classification problems[C]// Proc of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. La Jolla, USA, 2004: 73-78

[6] Zhang Yu, Wang Wen-jian, Guo Hu-sheng. An SVM accelerated training approach based on granular distribution[J]. Journal of Nanjing University(Natural Sciences), 2013, 49(5): 644-649(in Chinese)
张宇，王文剑，郭虎升. 基于粒分布的支持向量机加速训练方法[J]. 南京大学学报(自然科学版)，2013，49(5)：644-649

[7] Wang Wen-jian, Guo Hu-sheng. Granular support vector machine learning model[J]. Journal of Shanxi University(Natural Science Edition), 2009, 32(4): 535-540(in Chinese)
王文剑，郭虎升. 粒度支持向量机学习模型[J]. 山西大学学报(自然科学版)，2009，32(4)：535-540

[8] Zhang Wen-hao, Wang Wen-jian. A Kernel Granular Support Vector Machine Based on Association Rules [J]. Journal of Guangxi Normal University(Natural Sciences), 2009, 27(3): 89-92(in Chinese)
张文浩，王文剑. 一种基于关联规则的核粒度支持向量机[J]. 广西师范大学学报(自然科学版)，2009，27(3)：89-92

[9] Ertekin S, Huang Jian, Bottou L, et al. Learning on the border: active learning in imbalanced data classification [C]// Proceedings of the Sixteenth Conference ACM on Information and Knowledge Management. Lisbon, Portugal, 2007: 127-136

[10] Hu Zheng-ping, Gao Wen-tao. Training sample selection algorithm for SVM based on modified weighted condensed nearest neighbor and close-to-boundary criterion[J]. Journal of Yanshan University, 2010, 34(5): 421-426(in Chinese)
胡正平，高文涛. 基于改进加权压缩近邻与最近边界规则 SVM 训练样本约减选择算法[J]. 燕山大学学报，2010，34(5)：421-426

[11] de Sá J P M. Pattern recognition concepts, methods and applications [M]. Beijing: Tsinghua University Press, 2002: 21-22

[12] Du Shu-xin, Wu Tie-jun. Support vector machines for pattern recognition[J]. Journal of Zhejiang University (Technology Edition), 2003, 37(5): 521-528(in Chinese)
杜树新，吴铁军. 模式识别中的支持向量机方法[J]. 浙江大学学报(工学版)，2003，37(5)：521-528