第34卷 第16期 计 算 机 工 程 2008年8月
Vol.34 No.16 Computer Engineering August 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)16—0212—03 文献标识码：A 中图分类号：TP391

核子类凸包样本选择方法及其SVM应用

姜文瀚，周晓飞，杨静宇
(南京理工大学计算机科学与技术学院，南京 210094)

摘 要：提出一种基于核函数方法的类内训练样本选择方法——核子类凸包样本选择法，并将其用于支持向量机。该样本选择方法通过迭代方法，逐一选择了那些经映射后“距离已选样本”，并将其映射、生成“凸包最远的样本”。实验结果表明，该方法选择的少量样本使支持向量机获得了较高的识别比率，减少了存储需求，提高了分类速度。

关键词：样本选择；凸包；支持向量机；核函数；人脸识别

Kernel Subclass Convex Hull Sample Selection Method and Its Application on SVM

JIANG Wen-han, ZHOU Xiao-fei, YANG Jing-yu
(College of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094)

【Abstract】 A novel intra-class sample selection method named kernel subclass convex hull sample selection algorithm is proposed and used for SVM. The algorithm is an iterative procedure based on kernel trick. At each step, only one sample furthest to the convex hull spanned by chosen samples is picked out in the feature space. Experiments show that a significant amount of training data can be removed without sacrificing the performance of SVM, while the memory requirements and the computation time of the classifiers are reduced significantly.

【Key words】 sample selection; convex hull; support vector machine; kernel function; face recognition

1 概述

支持向量机(Support Vector Machine, SVM)以其坚实的理论基础和良好的泛化性能而被广泛应用于模式识别的诸多领域。然而求解凸二次规划问题带来的较高的计算代价却使得SVM在处理较大规模数据集的实际应用中面临困难，突出表现在空间占用和时间需求两个方面。SVM二次规划方程的核矩阵是 $N\times N$ (N 是两类训练样本数)的，标准二次规划问题(QP)求解算法的时间复杂度是$O(N^3)$[1]，如Matlab QP例程。显然，训练样本越多，所占用的存储空间越大，所耗费的计算处理时间越长。

为使SVM适应大数据集的应用，人们做了许多工作。总体而言，大概可以归纳为两个方面：一方面从改进优化过程入手，或采取分解的方法，或采用迭代的方法，将大的QP问题分成若干小的QP子问题加以解决，如Chunking方法，SMO，SVM^{light}以及Keerthi的快速迭代最近点方法[2]等。另一方面是遵循某种选择策略约简训练样本集。在保持SVM分类性能的前提下，样本筛选显然是降低计算代价的一个直接有效手段，可以有效地减少存储空间占用，节省计算时间。文献[3]利用随机方法为RSVM选择了部分训练样本，文献[4]使用K均值聚类方法从训练集中筛选边界样本。文献[1]依据近邻性质选择了那些位于边界附近的样本，文献[5]提出基于信任测度(confidence measure-based) 和 Hausdorff 距离 (Hausdorff distance-based)的两种不同样本选择方法。文献[6]将主动学习策略用于SVM的样本选择。

从SVM的几何解释可知，对于线性可分的训练样本集，SVM的最优分类超平面是两类训练样本凸包最近点对间连线的中垂面[2]。当SVM引入核函数后，最优分类超平面概念被推广到特征空间，其分类决策仍然由特征空间中各类映射样本的凸包所决定。凸包用其顶点或边缘点的凸组合就可以表达。凸包顶点的求解是一个典型的NP难题。为此，本文提出一种选择凸包边缘点的方法—核子类凸包样本选择方法。本文将核子类凸包样本选择方法用于SVM。在MIT-CBCL 人脸识别数据库training-synthetic子库上，该结合方法取得了较好的实验效果。

2 核子类凸包样本选择方法

核子类凸包样本选择方法是一种类内样本的选择方法。该方法针对一类训练样本集，迭代选择了那些经映射后距离已选样本映射凸包最远的样本(详见算法)。

已知训练样本集 $\boldsymbol{S}=\cup\boldsymbol{S}_i$， $i=1,2,\cdots,c$。 $\boldsymbol{S}_i$ 是第 i 类训练样本集合，包含有 n_i 个训练样本。假设存在某一映射关系 $\Phi:\boldsymbol{R}^n\to\boldsymbol{F}$，将原空间$\boldsymbol{R}^n$映射到某一高维特征空间 $\boldsymbol{F}$。那么，对于 $\boldsymbol{S}_i$，它在空间 $\boldsymbol{F}$ 中的映射集合为

$$\tilde{\boldsymbol{S}}_i=\{\Phi(\boldsymbol{x}_j)\mid \boldsymbol{x}_j\in\boldsymbol{S}_i, j=1,2,\cdots,n_i\}$$

由 $\tilde{\boldsymbol{S}}_i$ 中映射样本生成的类凸包定义为

$$co(\tilde{\boldsymbol{S}}_i)=\{\sum_{j=1}^{n_i}\alpha_j\Phi(\boldsymbol{x}_j)\mid\sum_{j=1}^{n_i}\alpha_j=1,\alpha_j\geq 0,\boldsymbol{x}_j\in\boldsymbol{S}_i\}$$

假设 $\boldsymbol{S}_i'$ 是 $\boldsymbol{S}_i$ 的选出样本集，且包含有 l 个已选择样本。$\boldsymbol{S}_i'$在空间 $\boldsymbol{F}$ 中的映射集合为 $\tilde{\boldsymbol{S}}_i'$，由 $\tilde{\boldsymbol{S}}_i'$ 中映射样本生成的子

作者简介：姜文瀚(1974－)，男，博士研究生，主研方向：模式识别，人工智能；周晓飞，博士研究生；杨静宇，博士生导师
收稿日期：2007-10-05 **E-mail：**e_wenhan@163.com

类凸包为

$$co(\tilde{\boldsymbol{S}}_i')=\{\sum_{j=1}^{l}\alpha_j\Phi(\boldsymbol{x}_j)\mid\sum_{j=1}^{l}\alpha_j=1,\alpha_j\geq 0,\ \boldsymbol{x}_j\in\boldsymbol{S}_i'\}$$

算法 (核子类凸包样本选择算法)

设第 i 类样本集 $\boldsymbol{S}_i$ 包含 n_i 个样本，已选样本子集 $\boldsymbol{S}_i'$，$\tilde{\boldsymbol{S}}_i'$ 是 $\boldsymbol{S}_i'$ 在特征空间的映射集合，l 是已选样本数。$k(\boldsymbol{x},\boldsymbol{y})$ 是核函数。

(1)初始化。设定拟选择样本个数 m，逼近误差界 ε，初始最大逼近误差 $e_{\max}=\inf$；初始选择集为

$$\boldsymbol{S}_i'=\{z_1,z_2\mid\ [z_1,z_2]=\arg\max_{\boldsymbol{x}_j,\boldsymbol{x}_k}\left\|\Phi(\boldsymbol{x}_j)-\Phi(\boldsymbol{x}_k)\right\|_2^2,\boldsymbol{x}_j,\boldsymbol{x}_k\in\boldsymbol{S}_i\}$$

其中，

$$\left\|\Phi(\boldsymbol{x}_j)-\Phi(\boldsymbol{x}_k)\right\|_2^2=(\Phi(\boldsymbol{x}_j)\cdot\Phi(\boldsymbol{x}_j))-2(\Phi(\boldsymbol{x}_j)\cdot\Phi(\boldsymbol{x}_k))+(\Phi(\boldsymbol{x}_k)\cdot\Phi(\boldsymbol{x}_k))=k(\boldsymbol{x}_j,\boldsymbol{x}_j)-2k(\boldsymbol{x}_j,\boldsymbol{x}_k)+k(\boldsymbol{x}_k,\boldsymbol{x}_k)$$

。

(2)如果选择集 $\boldsymbol{S}_i'$ 的样本个数 $l<m$，则对于 $\forall\boldsymbol{x}_p\in\boldsymbol{S}_i\backslash\boldsymbol{S}_i'$，计算 $d^2(\Phi(\boldsymbol{x}_p),co(\tilde{\boldsymbol{S}}_i'))$。令 $z_{l+1}=\arg\max_{\boldsymbol{x}_p\in\boldsymbol{S}_i\backslash\boldsymbol{S}_i'}d^2(\Phi(\boldsymbol{x}_p),co(\tilde{\boldsymbol{S}}_i'))$，$e_{\max}=d^2(z_{l+1},co(\tilde{\boldsymbol{S}}_i'))$；否则退出。

(3)如果 $e_{\max}>\varepsilon$，则 $\boldsymbol{S}_i'=\boldsymbol{S}_i'\cup z_{l+1}$；否则退出。

(4)返回(2)。

在算法中，步骤(2)所涉及的 $d^2(\Phi(\boldsymbol{x}_p),co(\tilde{\boldsymbol{S}}_i'))$ 的计算可归结为解如下优化方程：

$$d^2(\Phi(\boldsymbol{x}_p),co(\tilde{\boldsymbol{S}}_i'))=\min_{\boldsymbol{\alpha}}\left\|\Phi(\boldsymbol{x}_p)-\sum_{j=1}^{l}\alpha_j\Phi(z_j)\right\|_2^2=\min_{\boldsymbol{\alpha}}\left((\Phi(\boldsymbol{x}_p)\cdot\Phi(\boldsymbol{x}_p))-2\sum_{i=1}^{l}\alpha_i(\Phi(z_i)\cdot\Phi(\boldsymbol{x}_p))+\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j(\Phi(z_i)\cdot\Phi(z_j))\right)$$

$$\text{s.t. }\sum_{j=1}^{l}\alpha_j=1,\alpha_j\geq 0 \tag{1}$$

其中，$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_l)^{\mathrm{T}}$。

对于式(1)，不必知道具体映射函数 $\Phi(\cdot)$，而是通过核函数方法，将样本内积 $(\Phi(\cdot)\cdot\Phi(\cdot))$ 由核函数 $k(\cdot,\cdot)$ 来替代。式(1)可化为

$$d^2(\Phi(\boldsymbol{x}_p),co(\tilde{\boldsymbol{S}}_i'))=\min_{\boldsymbol{\alpha}}\left(k(\boldsymbol{x}_p,\boldsymbol{x}_p)-2\sum_{i=1}^{l}\alpha_i k(z_i,\boldsymbol{x}_p)+\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j k(z_i,z_j)\right)$$

$$\text{s.t. }\sum_{j=1}^{l}\alpha_j=1,\alpha_j\geq 0 \tag{2}$$

式(2)是一个凸二次规划式，该式涉及核矩阵大小(仅为 $l\times l$)。在实际应用中，每类选出集样本个数 l 一般要远小于该类别训练集样本数 n_i，因此，所需内存空间比较未经选样时的情况要少许多。

本文的核子类凸包样本选择方法的目的是要在特征空间中以较少的凸包边界样本尽可能地体现映射样本的类别分布，由选择样本集 $\boldsymbol{S}_i'$ 在特征空间的映射集合 $\tilde{\boldsymbol{S}}_i'$ 的子类凸包 $co(\tilde{\boldsymbol{S}}_i')$ 逼近类集 $\boldsymbol{S}_i$ 的所有映射样本的类凸包 $co(\tilde{\boldsymbol{S}}_i)$。当距离误差界 $\varepsilon=0$ 时，子类凸包 $co(\tilde{\boldsymbol{S}}_i')$ 实现了对类凸包 $co(\tilde{\boldsymbol{S}}_i)$ 的最佳逼近，$co(\tilde{\boldsymbol{S}}_i')$ 就是 $co(\tilde{\boldsymbol{S}}_i)$。在迭代过程中，到 $co(\tilde{\boldsymbol{S}}_i')$ 的距离为 0 的映射样本点由于已在凸包 $co(\tilde{\boldsymbol{S}}_i')$ 上，在以后迭代计算过程中，该点到 $co(\tilde{\boldsymbol{S}}_i')$ 的距离也必为 0，所以，可以省去，以减少计算量。

3 核子类凸包样本选择方法的 SVM 应用

假设给定两类问题的训练样本 $(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)$，$\boldsymbol{x}_i\in\boldsymbol{R}^n$，$y_i\in\{+1,-1\}$，$i=1,2,\cdots,N$，$y_i$ 表示类别标识。SVM的本质问题是求解如下的凸二次规划问题：

$$\max_{\boldsymbol{\alpha}}\sum_{j=1}^{N}\alpha_j-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_j\alpha_i\alpha_jk(\boldsymbol{x}_i,\boldsymbol{x}_j)$$
$$\text{s.t. }\sum_{i=1}^{N}y_i\alpha_i=0\quad 0\leq\alpha_i\leq C,i=1,2,\cdots,N \tag{3}$$

本文首先采用核子类凸包样本选择方法，针对各类训练集进行类内选样，然后用各类选出的样本训练 SVM。核子类凸包样本选择方法的核函数及参数与 SVM 的相同。假设选样后的各类训练样本个数为 $l(2l<<N)$，则优化问题式(3)中的核矩阵的大小将为 $2l\times 2l$，这较未经选样的 SVM 所涉及的原训练集核矩阵 $N\times N$ 要小的多。样本集的缩小大大降低了 SVM 的内存需求，减少了不必要的冗余计算，从而有效地保证了 SVM 的顺利执行。

4 人脸识别实验及分析

为验证核子类凸包样本选择方法用于SVM的有效性，本文在MIT-CBCL人脸识别数据库[7]的training-synthetic子库上对采用径向基核函数的核子类凸包样本选择方法的SVM进行了验证测试。该子库包含 3D形态模型合成的姿态和光照变化的 10 个人的 3 240 幅标准人脸图像，每人 324 幅。PGM格式，分辨率 200×200。其中人脸姿态变化为：水平向左旋转 0°~32°，以 4°为增量。光照角度变化为：以头部为中心，水平右向 15°~90°，以 15°为增量；竖直仰角 0°~75°，以 15°为增量。图 1 是该子库中一个人的部分示例图像。

图 1 MIT-CBCL 人脸识别 training-synthetic 库示例

本节实验把该子库中的人脸图像按照光照仰角分为两个子集A_1和A_2，分别包括仰角为{0°, 30°, 60°}和{15°, 45°, 75°}的图像。两个子集各包括 10 个人的 1 620 幅图像，每人 162 幅。实验将全部图像转换为JPG格式，并双 3 次插值缩为 16×16 大小。

实验在Pentium IV 2.8 GHz CPU, 256 MB内存的PC机上执行。SVM的多类分类采用自底向上二叉树结构[8]来分解实现。样本选择以选择样本个数作为终止条件。SVM参数 $C=\infty$。实验所涉及的优化过程均由Matlab的优化工具实现。

实验首先将采用径向基核函数的核子类凸包样本选择方法的径向基核SVM与采用随机样本选择方法的径向基核SVM进行比较。径向基核函数的尺度参数 $\sigma=6$。实验以A_1为训练集，从中选择样本训练，并分成两种情况进行测试。一种情况是在训练集A_1自身上进行测试，目的是衡量选出样本对训练集类别的表达能力。另一种情况是对A_2进行测试，目的是反映选择样本对SVM分类器性能的影响。针对不同选样个数(逐渐增加选样个数)，两种选样方法的SVM在A_1和A_2上的测试识别率见表 1。选样率=每类选样数/162×100%，测试识别率＝正确识别样本数/1 620×100%；随机选样的识别率为 10 次随机选样测试的平均识别率。从实验数据可以看到，在以 $\boldsymbol{A}_1$为训练集的情况下，随着选择样本数量的增加，采用核子类凸包样本选择方法的SVM的泛化能力提升很快，当选样

率为 3.7%，即各类别选择 6 个样本时，A_1和A_2的测试识别率均已能够达到 100%。与之相对的采用随机选样的SVM仅分别获得了 96.52%和 95.20%的平均识别率。

表 1　本文方法和随机选样方法下的径向基核 SVM 实验结果

选样数/类	选样率/(%)	测试识别率			
		核子类凸包样本选择+SVM		随机选样+SVM	
		测试A_1/(%)	测试A_2/(%)	测试A_1/(%)	测试A_2/(%)
2	1.2	92.22	89.32	62.65	59.87
3	1.9	98.33	96.98	83.20	82.06
4	2.5	99.01	97.35	90.24	87.81
5	3.1	98.95	97.16	95.12	93.41
6	3.7	100.00	100.00	96.52	95.20
7	4.3	100.00	100.00	98.21	97.72
8	4.9	100.00	100.00	99.04	98.78

实验还将本文采用径向基核函数核子类凸包样本选择方法的径向基核 SVM(选样算法和分类器的尺度参数 σ 均为 6)与不经过任何样本选择的径向基核 SVM(尺度参数 σ 为 0.3)在执行时间上进行了测试比较，结果见表 2。

表 2　未选样和本文选样的径向基核 SVM 的实验比较

实验方法	选样数/类	识别率/(%)	选样时间/s	测试时间/s	合计时间/s
SVM	未选	100	-	363.656	363.656
本文选样法+SVM	6	100	18.031	7.891	25.922

在表 2 中，本文核子类凸包样本选择方法下的 SVM 每类选择 6 个样本，正确识别率 100%，用时 25.922 s。较相同识别率的未经样本选择的 SVM 在测试时间上有了明显减少。

表 1 和表 2 的实验数据充分说明，本文的核子类凸包样本选择方法在保证 SVM 具有良好泛化性能的前提下，能够有效地实现 SVM 训练集样本约简，降低存储需求，并且加快分类速度。

5　结束语

本文提出核子类凸包样本选择方法，并将其用于 SVM。核子类凸包样本选择方法是一种类内样本选择方法，针对每一类训练样本，迭代选择了那些经映射后距离已选样本映射生成凸包最远的样本。在 MIT-CBCL 人脸识别数据库的 training-synthetic 子库上，与采用随机选样策略的 SVM 和未进行样本选择的 SVM 相比，本文选样方法下的 SVM 表现了训练样本少、泛化能力强、计算速度快的优点。

参考文献

[1] Shin H, Cho S. Neighborhood Property Based Pattern Selection for Support Vector Machines[J]. Neural Computation, 2007, 19(3): 816-855.

[2] Keerthi S S, Shevade S K, Bhattacharyya C. A Fast Iterative Nearest Point Algorithm for Support Vector Machine Classifier Design[J]. IEEE Transactions on Neural Networks, 2000, 11(1): 124-136.

[3] Lee Y, Mangasarian O L. RSVM: Reduced Support Vector Machines[C]//Proc. of the SIAM International Conference on Data Mining. San Jose, CA: [s. n.], 2001.

[4] Almeida M B, Braga A P, Braga J P. Svm-km: Speeding Svms Learning with a Priori Cluster Selection and K-means[C]//Proc. of the 6th Brazilian Symposium on Neural Networks. Rio de Janeiro, Braz: [s. n.], 2000: 162-167.

[5] Wang Jigang. Training Data Selection for Support Vector Machines[J]. Lecture Notes in Computer Science, 2005, 3610: 554-564.

[6] Schohn G, Cohn D. Less is More: Active Learning with Support Vector Machines[C]//Proceedings of the 17th International Conference on Machine Learning. [S. l.]: IEEE Press, 2000: 839-846.

[7] Weyrauch B, Huang J, Heisele B, et al. Component-based Face Recognition with 3D Morphable Models[C]//Proc. of the 1st IEEE Workshop on Face Processing in Video. Washington, D. C., USA: [s. n.], 2004.

[8] Guo G D, Li S Z. Support Vector Machines for Face Recognition[J]. Image and Vision Computing. 2001, 19(9/10): 631-638.

(上接第 211 页)

的兴趣模型，在发生第 1 次兴趣漂移前，随着时间窗口的增大，主题分类更加精确，分类错误率也逐渐降低。当发生兴趣漂移时，错误率显著升高，加入了时间窗优化算法的兴趣模型在发现错误率显著升高后，调整时间窗使错误率逐渐降低。不使用优化时间窗的兴趣模型，在发生第 1 次兴趣漂移前，随着时间窗口的增大，分类错误率逐渐降低。当发生兴趣漂移后，兴趣模型描述用户兴趣越来越不准确，所以分类错误率一直逐渐升高。经过多次试验证明，优化时间窗算法适合于常用的分类算法，能帮助兴趣模型提供更加准确的兴趣信息。

4　结束语

本文提出一种基于优化时间窗的用户兴趣漂移方法，借助分类错误率跟踪用户兴趣变化，然后通过改进的时间窗算法调节时间窗大小。该方法不但能准确地发现用户兴趣的变化，而且还能通过调节时间窗设置精确的窗口大小。基于优化时间窗的方法能够比较准确地描述用户兴趣，具有较高的效率。将来的工作中，还可以对算法进行其他尝试，如采用多窗口跟踪用户兴趣，通过设置不同的 K 值，对不同的用户设置不同的遗忘速度。

参考文献

[1] Klinkenberg R. Learning Drifting Concepts: Example Selection vs. Example Weighting[J]. Intelligent Data Analysis, 2004, 8(3): 281-300.

[2] Koychev I, Schwab I. Adaptation to Drifting User's Intersects[C]// Proc. of the Workshop on Machine Learning in New Information Age. Barcelona, Spain: [s. n.], 2000.

[3] Maloof M A, Ryszard S. Selecting Examples for Partial Memory Learning[J]. Machine Learning, 2000, 41(1): 27-52.

[4] Widmer G, Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts[J]. Machine Learning, 1996, 23(1): 69-101.