

核仿射子空间最近点分类算法

周晓飞, 姜文瀚, 杨静宇

(南京理工大学计算机科学与技术学院, 南京 210094)

摘要: 受支持向量机的几何解释和最近点问题启发, 提出一种新型的模式分类算法——核仿射子空间最近点分类算法。该算法在核空间中, 将支持向量机几何模型中的最近点搜索区域由 2 类训练特征集凸包推广到 2 类特征样本各自生成的仿射子空间, 以仿射子空间作为特征样本分布的粗略估计, 通过仿射子空间中的最近的 2 个点构造平分仿射子空间间隔的最优分类超平面。该算法在 ORL 人脸识别数据库上的比较实验中取得了较好的识别效果。

关键词: 模式分类; 核函数; 支持向量机; 核仿射子空间最近点

Kernel Affine Subspace Nearest Points Classification Algorithm

ZHOU Xiao-fei, JIANG Wen-han, YANG Jing-yu

(School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094)

【Abstract】 A novel pattern recognition algorithm called Kernel Affine Subspace Nearest Points(KASNP) classification is presented. Inspired by the geometrical explanation of Support Vector Machine(SVM) that the optimal separating plane bisects the closest points within two class convex hulls, KASNP algorithm expands the searching areas of the closest points from the convex hulls to their corresponding class affine subspaces in kernel space. The affine subspaces are taken as the rough estimations of the class feature sample distributions, and their closest points are found. The hyperplane to separate the affine subspaces with the maximal margins is constructed, which is the perpendicular bisector of the line segment joining the two closest points. The test experiments compared with the Nearest Neighbor(1-NN) classifier and SVM on the ORL face recognition database show good performance of this algorithm.

【Key words】 pattern classification; kernel function; Support Vector Machine(SVM); Kernel Affine Subspace Nearest Points(KASNP)

1 概述

文献[1-2]从凸包角度解释了支持向量机(Support Vector Machine, SVM)^[3]的几何本质。SVM 在特征空间所求取的最优分类超平面实质是以 2 类训练样本凸包最近点连线方向为法向量、且过最近点连线中点的超平面。因此, 该最优分类超平面也最大间隔地分离了 2 类训练样本的凸包。从 SVM 的几何解释分析来看, SVM 隐含地将训练集扩展为各类训练样本的凸包, 由凸包作为样本分布的参照模型, 并构造最大间隔分离凸包的分类超平面。在实际应用中, 许多高维自然数据往往都分布在高维空间的低维流形上, 而比较简单直观的低维流形就是训练样本所在的最小线性流形 - 仿射子空间。本文将仿射子空间替代 SVM 几何模型中的凸包, 提出一种新的基于核函数方法的模式分类算法——核仿射子空间最近点分类算法(Kernel Affine Subspace Nearest Points, KASNP)。

2 支持向量机的几何解释

SVM 方法所寻找的最优分类超平面实质上就是垂直平分 2 类训练样本凸包最近点对连线的超平面^[1-3]。图 1 以二维空间样本为例, 示意了 SVM 的几何本质。图 1 中凸包边缘用虚线表示, H 是最优分类超平面。SVM 算法所要寻找的最佳投影方向 w 正是 2 类样本凸包最近点(图 1 中 c, d)连线方向; 支持向量机中所谓的最大间隔(如图 1 中支撑面 H_1 和 H_2 间的距离 $2/\|w\|$)就是凸包间隔(最近点 c, d 间的距离)。因此, SVM 的最大间隔问题可被转化为凸包最近点的问题。下面通过求取凸包最近点来构造最优分类超平面的 SVM 几何方

法, 文献[4]称其为平分最近点法。

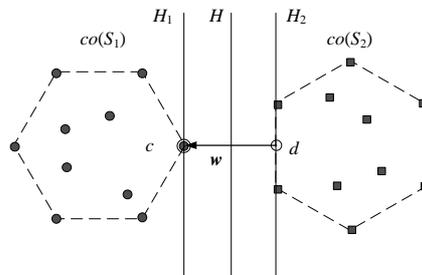


图 1 SVM 的几何解释

假设 R^n 中有 2 类训练样本 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$, $y_i \in \{+1, -1\}$ 表示类别标识。2 类训练集分别可表示为 $S_1 = \{x_i | y_i = +1\}$ 和 $S_2 = \{x_i | y_i = -1\}$ 。 S_1 和 S_2 的凸包分别为

$$co(S_1) = \left\{ \sum_{y_i=+1} \alpha_i x_i \mid \sum_{y_i=+1} \alpha_i = 1, 0 \leq \alpha_i \leq 1 \right\} \quad (1)$$

$$co(S_2) = \left\{ \sum_{y_i=-1} \alpha_i x_i \mid \sum_{y_i=-1} \alpha_i = 1, 0 \leq \alpha_i \leq 1 \right\} \quad (2)$$

假定 2 类训练样本凸包不相交(训练集线性可分), 求解 2 个凸包上最近点的优化问题为

基金项目: 国家自然科学基金资助项目(60472060, 60632050)

作者简介: 周晓飞(1978 -), 女, 博士研究生, 主研方向: 模式识别, 人工智能; 姜文瀚, 博士研究生; 杨静宇, 教授、博士生导师

收稿日期: 2007-09-15 **E-mail:** zhouxf@njjust.edu.cn

$$\begin{aligned} \min_a \quad & \left\| \sum_{y_i=+1} \alpha_i x_i - \sum_{y_i=-1} \alpha_i x_i \right\|^2 \\ \text{s.t.} \quad & \sum_{y_i=+1} \alpha_i = 1, \sum_{y_i=-1} \alpha_i = 1, \alpha_i \geq 0, i=1,2,\dots,l \end{aligned} \quad (3)$$

式(3)为凸二次规划问题,假设通过优化求得的最优解为 $a^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_l^*)$, 那么 2 类凸包的最近点对为

$$c = \sum_{y_i=+1} \alpha_i^* x_i, \quad d = \sum_{y_i=-1} \alpha_i^* x_i \quad (4)$$

构造垂直平分线段 \overline{cd} 的分类超平面 $w \cdot x + b = 0$ 。其中,法向量 $w = c - d$; 由于该超平面过最近点连线的中点 $p = \frac{1}{2}(c + d)$, 因此 $b = -w \cdot p = -\frac{1}{2}(c - d)(c + d)$ 。决策函数则为 $f(x) = \text{sgn}(w \cdot x + b)$ 。

同样,对于非线性问题,非线性 SVM 的几何本质可以解释为:通过核函数方法在核空间训练构造平分特征样本凸包间隔的最优分类超平面。

3 核仿射子空间最近点分类算法

受 SVM 的几何解释启发,本文在核空间将 SVM 几何方法中的最近点搜索区域由凸包推广到仿射子空间,并以平分各类训练集的仿射子空间的间隔为目的来构造分类超平面,进而得到一种新型的分类方法——核仿射子空间最近点分类算法。与 SVM 最近点问题的凸包不相交前提类似,核仿射子空间最近点分类算法执行的前提条件是核空间 2 个仿射子空间不相交。

假设 2 类训练样本 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$, $x_i \in R^n$, $y_i \in \{+1, -1\}$ 为类别标识。2 类训练集分别为 $S_1 = \{x_i | y_i = +1\}$ 和 $S_2 = \{x_i | y_i = -1\}$ 。 $k(x, y) = \Phi(x)^T \Phi(y)$ 为核函数,其中,特征映射 $\Phi: R^n \rightarrow F$ 将训练集 S_1 和 S_2 所在的输入空间 R^n 映射到特征空间 F 。则 S_1 和 S_2 在 F 空间中映射集分别为

$$\tilde{S}_1 = \{\Phi(x_i) | \Phi: R^n \mapsto F, x_i \in R^n, y_i = +1\}$$

$$\tilde{S}_2 = \{\Phi(x_i) | \Phi: R^n \mapsto F, x_i \in R^n, y_i = -1\}$$

那么, \tilde{S}_1 和 \tilde{S}_2 张成的 2 个仿射子空间分别为

$$H(\tilde{S}_1) = \left\{ \sum_{y_i=+1} \alpha_i \Phi(x_i) \mid \sum_{y_i=+1} \alpha_i = 1 \right\} \quad (5)$$

$$H(\tilde{S}_2) = \left\{ \sum_{y_i=-1} \alpha_i \Phi(x_i) \mid \sum_{y_i=-1} \alpha_i = 1 \right\} \quad (6)$$

在特征空间 F 中构造仿射子空间最近点问题的优化函数

$$\begin{aligned} \min_a \quad & \left\| \sum_{y_i=+1} \alpha_i \Phi(x_i) - \sum_{y_i=-1} \alpha_i \Phi(x_i) \right\|^2 \\ \text{s.t.} \quad & \sum_{y_i=+1} \alpha_i = 1, \sum_{y_i=-1} \alpha_i = 1 \end{aligned} \quad (7)$$

其中, $a = (\alpha_1, \alpha_2, \dots, \alpha_l)^T$ 。

式(7)可化为只与训练样本核函数有关的优化问题:

$$\begin{aligned} \min_a \quad & \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j k(x_i, x_j) \\ \text{s.t.} \quad & \sum_{y_i=+1} \alpha_i = 1, \sum_{y_i=-1} \alpha_i = 1 \end{aligned} \quad (8)$$

求得优化式(8)最优解 $a^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_l^*)^T$ 后,可以求出仿射子空间的 2 个最近点对:

$$c = \sum_{y_i=+1} \alpha_i^* \Phi(x_i), \quad d = \sum_{y_i=-1} \alpha_i^* \Phi(x_i)$$

构造垂直平分最近点对线段 \overline{cd} 的分类超平面 $w \cdot \Phi(x_i) + b = 0$ 。

$$w = c - d = \sum_{y_i=+1} \alpha_i^* \Phi(x_i) - \sum_{y_i=-1} \alpha_i^* \Phi(x_i) = \sum_{i=1}^l y_i \alpha_i^* \Phi(x_i)$$

最近点连线的中点 p :

$$p = \frac{1}{2}(c + d) = \frac{1}{2} \left(\sum_{y_i=+1} \alpha_i^* \Phi(x_i) + \sum_{y_i=-1} \alpha_i^* \Phi(x_i) \right) = \frac{1}{2} \sum_{i=1}^l \alpha_i^* \Phi(x_i)$$

由于分类超平面过 p , 有 $w \cdot p + b = 0$, 可得

$$b = -w \cdot p = -\frac{1}{2}(c - d) \cdot (c + d) = -\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i \alpha_i^* \alpha_j^* k(x_i, x_j)$$

于是判决函数为

$$f(x) = \text{sgn} \left(\sum_{i=1}^l y_i \alpha_i^* k(x, x_i) - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i \alpha_i^* \alpha_j^* k(x_i, x_j) \right)$$

对于测试样本 x , 如果 $f(x) > 0$, 则 x 属于正类, 否则属于负类。

从几何角度出发,本文提出的核仿射子空间分类算法与非线性 SVM 的主要差异在于:非线性 SVM 是在核空间训练构造平分凸包间隔的最优线性决策界,而本文算法是在核空间训练构造平分仿射子空间间隔的分类超平面。图 2 在三维空间示意了 SVM 凸包最近点问题与仿射子空间最近点问题的差别。

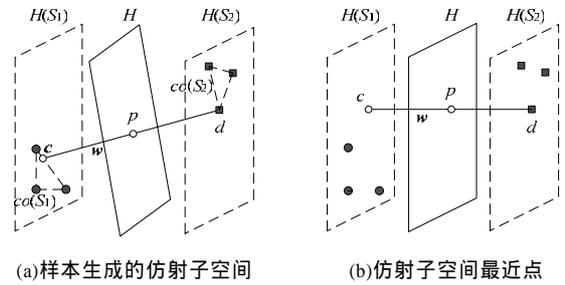


图 2 本文方法与 SVM 的几何原理比较

图 2 中的 $H(S_1)$ 和 $H(S_2)$ 分别表示 2 类样本生成的仿射子空间, $co(S_1)$ 和 $co(S_2)$ 为 2 类样本凸包。图 2(a)中的 c, d 为凸包最近点,图 2(b)中的 c, d 为仿射子空间最近点。图中标识为 H 的分类面过 p 点,且法向量为 dc 。

本文的核仿射子空间最近点分类算法也是 2 类问题分类器。当其用于处理多类问题时,可以采用自底向上二叉树结构方法^[5]将多类分类问题转化为多个 2 类问题来解决。每一个 2 类核仿射子空间最近点分类器则作为树的判别节点。

4 人脸识别实验分析

实验选择在 ORL 标准人脸图像库上进行。ORL 标准人脸库由 40 个人的 400 幅灰度图像构成,每人 10 幅,图像格式为 PGM,分辨率是 92×112 。图 3 是该库一个人的人脸图像示例。实验将全部图像转换为 JPG 格式,并双三次插值缩至 16×16 。



图 3 ORL 人脸数据库图像示例

实验将核仿射子空间最近点分类算法同最近邻分类器(one Nearest Neighbor, 1-NN)和支持向量机进行测试比较。SVM 和 KASNP 分别采用线性(linear)核 $k(x, y) = (x \cdot y)$ 和径向基(rbf)核 $k(x, y) = \exp(-0.5 \|x - y\|^2 / \sigma^2)$ 。对于多类问题,对应 SVM 和 KASNP 方法的实验采用自底向上二叉树结构方法转化为

多个 2 类问题来处理^[5]。

实验共分 10 组进行,每组的训练集由 200 幅图像及其镜像图像构成,剩余 200 幅用于测试。10 组实验(实验序号为 1~10)训练集各类样本的相应标识为 {1,2,3,4,5}, {2,3,4,5,6}, ..., {6,7,8,9,10}, {7,8,9,10,1}, {8,9,10,1,2}, ..., {10,1,2,3,4}。实验中 SVM 的径向基核参数 $\sigma = 6$,KASNP 的径向基核参数 $\sigma = 2$,各 SVM 的惩罚系数 $C = \infty$ 。实验在 P4 2.8 GHz, 256 MB 内存的 PC 上执行,SVM 的优化求解采用 Matlab quadprog 例程实现。10 组实验的各分类器的错误识别样本个数见表 1,表中还统计了 10 组实验的合计误识样本总数。

表 1 ORL 人脸数据库上的测试集错误识别样本个数比较

实验序号	错误识别个数				
	1-NN	SVM(linear)	SVM(rbf)	KASNP(linear)	KASNP(rbf)
1	14	15	15	12	13
2	10	10	11	8	6
3	7	10	9	9	10
4	11	3	3	5	4
5	11	4	4	5	4
6	15	12	12	9	8
7	8	6	5	4	2
8	6	6	6	4	2
9	9	4	4	4	3
10	12	8	9	7	8
合计误识数	103	78	78	67	60

在 10 组实验中,径向基核 KASNP 算法有 6 组实验的错误样本数最少,1-NN, SVM(linear), SVM(rbf)和 KASNP(linear)分别有 1, 2, 2, 2 组实验错误数最少。从合计错误识别样本数来看,1-NN 的合计错误样本数最多,为 103;SVM(linear)和 SVM(rbf)合计错误样本数相同,为 78;线性核 KASNP 合计错误样本数少于前 3 个分类器,为 67;而径向基核 KASNP 的合计错误数较线性核 KASNP 还要少 7 个。

另外,与其他的核学习方法(如 SVM)一样,KASNP 的性能同样受到核函数的选择及其参数设置的影响。而核函数及其参数的选择问题仍然是核机器学习研究中的难题,目前尚没有成熟有效的解决方法。因此,实验中 KASNP 和 SVM 的径向基核函数的尺度参数是凭借经验设定的,最终选取了使平均识别效果最佳的参数值。

由于核仿射子空间最近点问题的优化函数式不含约束项 $\alpha_i = 0$,较标准 SVM 的凸二次规划问题求解容易得多,因此在核函数相同的情况下,KASNP 训练速度比标准 SVM 快,

有效节省了执行时间。为证实这一点,实验还记录了 10 组实验中 SVM 和 KASNP 方法的平均执行时间(包括训练和测试时间),见表 2。

表 2 KASNP 与 SVM 平均执行时间比较 s

线性核情况		径向基核情况	
SVM(linear)	KASNP(linear)	SVM(rbf)	KASNP(rbf)
31.933	20.314	43.022	30.962

从表 2 可以看到,无论是采用线性核函数,还是采用径向基核函数,本文 KASNP 算法的平均执行速度均比 SVM 快。线性核情况下,KASNP 的平均时间比 SVM 少 11.619 s;径向基核情况下,KASNP 的平均时间比 SVM 少 12.06 s。

5 结束语

本文提出了核仿射子空间最近点分类算法。该分类算法首先通过核函数方法将样本映射到高维特征空间,然后在高维特征空间里构造最大间隔分离 2 类特征样本仿射子空间的分类超平面。由于仿射子空间的最近点优化函数式不含约束项 $\alpha_i = 0$,因此其求解比 SVM 容易得多。在 ORL 人脸识别数据库上,核仿射子空间最近点分类算法不但取得了高于最近邻分类器、SVM 分类器的平均识别率,而且与 Matlab 优化工具箱实现的 SVM 分类器相比,具有更快的分类速度。

参考文献

- [1] Keerthi S S, Shevade S K, Bhattacharyya C, et al. A Fast Iterative Nearest Point Algorithm for Support Vector Machine Classifier Design[J]. IEEE Transactions on Neural Networks, 2000, 11(1): 124-136.
- [2] Bennett K P, Bredensteiner E J. Duality and Geometry in SVM Classifiers[C]//Proceedings of the 17th International Conference on Machine Learning. San Francisco, California, USA: Morgan Kaufmann, 2000: 57-64.
- [3] Vapnik V N. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000.
- [4] 邓乃扬,田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京:科学出版社,2004.
- [5] Guo Guodong, Sann Z L. Support Vector Machines for Face Recognition[J]. Image and Vision Computing. 2001, 19(9/10): 631-638.

(上接第 22 页)

4 结束语

本文介绍了一个基于编辑距离和多种后处理的生物实体名识别方法:先使用全称缩写对识别算法扩充词典,再通过编辑距离算法提高召回率,并采用多种后处理方法进一步提高性能。试验结果表明,即使基于内部词典,该方法也能大幅度提高识别的性能。在下一步的工作中,将引入规模更大的外部词典考察对实体识别效果的影响。

参考文献

- [1] Lee K J, Hwang Y S, Rim H C. Two-phase Biomedical NE Recognition Based on SVMs[C]//Proceedings of the ACM 2003 Workshop on Natural Language Processing in Biomedicine. Sapporo, Japan: [s. n.], 2003: 33-40.
- [2] Zhou Gudong, Su Jian. Exploring Deep Knowledge Resources in

- Biomedical Name Recognition[C]//Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Geneva, Switzerland: [s. n.], 2004: 96-99.
- [3] Settles B. Biomedical Named Entity Recognition Using Conditional Random Fields and Novel Feature Sets[C]//Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Geneva, Switzerland: [s. n.], 2004: 104-107.
- [4] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proceedings of the International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, 2001: 282-289.
- [5] Navarro G. A Guided Tour to Approximate String Matching[J]. ACM Computing Surveys, 2001, 33(1): 31-88.