基于核变换的高性能支持向量机分类算法

毕德学 于德敏 许增朴
(天津科技大学机械工程学院, 天津 300222)

摘 要 由于传统的支持向量机(SVM)算法的核函数没有考虑训练数据自身的特点,因而相对于具体的问题来说,往往不是最优的。为了获得最优的分类结果,提出了一种基于核变换思想的支持向量机分类方法。该方法首先根据训练样本的类属信息,通过对初始核进行线性变换来间接地达到改进输入空间到输出空间的映射函数的目的,同时利用变换后的核函数来求解分类数据特征空间的超平面方程。仿真和实验结果表明,采用此方法,不仅可以提高系统的分类性能和降低噪声的干扰,而且可以增强分类结果的鲁棒性。

关键词 支持向量机 核变换 特征空间
中图法分类号:TP181 **文献标识码**:A **文章编号**:1006-8961(2008)10-1845-04

High Performance SVM Classification Based on Kernel Transformation

BI De-xue, YU De-min, XU Zeng-pu
(*College of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin* 300222)

Abstract Kernel based Support Vector Machine (SVM) does not consider inner property of training data, so classification results are usually not in optimum condition. In this paper we present a new SVM classification algorithm. The proposed method alters the kernel based on the class information of the training data, with input vectors being classified by this transformed kernel. The described algorithm can improve performance of mapping function indirectly. Simulation and experiments validate that it can improve classification performance and robustness, and reduce noise.

Keywords SVM, kernel transformation, feature space

1 引 言

支持向量机算法可以通过将低维的输入空间数据映射到一个高维的特征空间,并可利用核函数在此特征空间中来寻找最优的分类超平面,因此,如何创建这个从低维的输入空间到高维的特征空间函数的映射,其定义的核函数对分类性能来说至关重要。文献[1]提出了基于黎曼几何的通过增大映射空间分类面附近的空间分辨力来提高系统分类性能的方法,并用改进的高斯核函数做了实验,取得了较好的效果。但该方法需要仔细选好空间分辨率权重函数及核改进前的支持向量,过程较为繁琐,该文也没有明确从理论上证明系统的泛化能力是否有提高。文献[2]、[3]则采用基于距离的方法进行分类,并提高了分类的精度和抗噪能力。

由于支持向量机(SVM)的最优超平面对噪声比较敏感,因此研究者提出了各种消除噪声影响的方法。文献[4]提出了一种基于归一化核函数的分类方法,并通过归一化映射空间来减少噪声影响,以提高系统的分类性能。另外一种常用的方法是模糊支持向量机方法[5,6],由于该方法对不同的样本采用不同的惩罚系数,使得构造目标函数时,不同的样本有不同的贡献,并对含有噪声与野值的样本赋予较小的

基金项目:国家自然科学基金项目(60675046)
收稿日期:2008-06-20;**改回日期**:2008-07-15
第一作者简介:毕德学(1969~),男。副教授。2004 年获香港城市大学博士学位。主要研究方向为机器智能,机器视觉。
E-mail:dexue@tust.edu.cn

权值,从而在一定程度上可消除噪声与野值的影响。文献[7]、[8]提出了基于样本紧密度的隶属度确定方法,但是该方法没有考虑不同类的类中心对确定该隶属度的影响;另外,此方法是基于原始空间进行的,当把原始空间中的点映射到高维特征空间时,原始空间的点将在高维特征空间重新分布。

根据文献[9]有关理想核的定义和特性,本文提出了一种简单有效的基于核变换的支持向量机分类算法[10,11],该算法首先根据输入数据的类属性来对核函数进行平移变换,使不同类间的数据的距离增大,即相似性减小;然后使相同类间的数据距离减小,即相似性增大,从而创造出了一个新的核函数矩阵。该方法由于无需求解优化的映射函数,而是通过直接对初始核函数进行线性变换,间接地可达到对输入空间到输出空间的映射函数进行改进的目的,因此可以大大提高系统的抗噪声能力和识别精度。

2 基于线性变换的归一化核

传统支持向量机的分类性能对核函数的依赖程度非常高,即当选定了核函数,就选定了一个从输入空间到特征空间的映射。由于该映射完全没有考虑训练数据自身的特点,这相对于具体问题来说,往往不是最优的,因此有些情况下,其分类精度不够高,且对噪声比较敏感、数值输出不够稳定、泛化能力不够强。为了克服这类问题,本文提出了一种基于核变换的支持向量机分类方法。

定理 1 归一化核经过如下的线性变换后,仍然是归一化核:

$$\begin{cases}\hat{k}_1(\boldsymbol{x}_i,\boldsymbol{x}_j) = \lambda_1 k_N(\boldsymbol{x}_i,\boldsymbol{x}_j) + 1 - \lambda_1 & y_i = y_j \\ \hat{k}_2(\boldsymbol{x}_i,\boldsymbol{x}_j) = \lambda_1 k_N(\boldsymbol{x}_i,\boldsymbol{x}_j) & y_i \neq y_j \end{cases} \quad (1)$$

$$0 \leqslant \lambda_1 \leqslant 1$$

证明 变换后的核函数 $\hat{k}$ 可以写成如下方程:

$$\hat{k} = \lambda_1 k_N + (1 - \lambda_1) k^* \quad (2)$$

式中,k^* 表示理想的核函数,即当两个输入向量是同类时,它们的相似度为 1,反之,它们的相似度为 0。文献[9]已经证明与理想的核函数对应的 Gram 矩阵是正定(半正定)的,由于输入的 k_N 是归一化核函数,因此由它组成的 Gram 矩阵是正定(半正定)的,根据式(1)的条件 $1 \geqslant 1 - \lambda_1 \geqslant 0$ 可知,变换后的核函数 $\hat{k}$ 是由两个有效核函数的非负加权的和组成。根据核的封闭特性[12]可知,有效核经过加法与乘法(系数为非负)运算后仍然是有效核,因此,变换后的核函数 $\hat{k}$ 仍然是有效的核函数。又由于 k_N 是归一化核函数,即 $k_N(\boldsymbol{x}_i,\boldsymbol{x}_i) = 1$,因此 $\hat{k}(\boldsymbol{x}_i,\boldsymbol{x}_i) = 1$,即变换后的核函数 $\hat{k}$ 是归一化的。由此定理得证。

3 基于归一化核变换的分类算法

支持向量机算法在采用归一化核变换(式(1))进行分类时,由于在整个输入空间上的核是非唯一的,因此用传统的算法无法直接求解。下面,本文将推导出能采用不同核(式(1))的支持向量机分类算法。

根据定理(1),由于变换后的核是有效的,因此存在一个映射 $\varphi(\boldsymbol{x})$,满足变换后的核函数(式(1))。设原来的未变换的归一化核函数的映射是 $\phi(\boldsymbol{x})$,并给定输入训练样本集合 $\boldsymbol{X}: = \{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\} \subset \boldsymbol{\chi}$,以及与每个训练样本对应的分类输出 $\boldsymbol{Y}: = \{y_1,y_2,\cdots,y_n\} \subset \{-1,+1\}$,则基于支持向量机实现分类优化的目标是[11]:

$$\min_{\pi,b,\zeta} \frac{1}{2}\|\pi\|^2 + c\sum_{i=1}^{n}\zeta_i \quad (3)$$

$$y_i(\langle \pi,\varphi(\boldsymbol{x}_i)\rangle + b) \geqslant 1 - \zeta_i \qquad i = 1,2,\cdots,n$$

$$\zeta_i \geqslant 0 \qquad i = 1,2,\cdots,n \quad (4)$$

式中,$\varphi(\boldsymbol{x})$ 是一个经过式(1)核变换后的低维空间到高维空间的映射。π 表示优化目标的复杂度,ζ_i 表示样本的训练误差,c 表示权重系数。其对应的拉格朗日函数为

$$L(\pi,b,\zeta,\alpha) = \frac{1}{2}\|\pi\|^2 + c\sum_{i=1}^{n}\zeta_i - \sum_{i=1}^{n}\alpha_i(y_i(\langle \pi,\varphi(\boldsymbol{x}_i)\rangle + b - 1 - \zeta_i)) - \sum_{i=1}^{n} r_i\zeta_i \quad (5)$$

这里,$\alpha_i \geqslant 0, r_i \geqslant 0$,通过求对应于 π,b,ζ 的偏导,即可得到下面的对偶目标函数:

$$L(\pi,b,\zeta,\alpha,r) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{\substack{i,j=1 \\ y_i=y_j}}^{n} y_i y_j \alpha_i \alpha_j \hat{k}_1(\boldsymbol{x}_i,\boldsymbol{x}_j) - \frac{1}{2}\sum_{\substack{i,j=1 \\ y_i\neq y_j}}^{n} y_i y_j \alpha_i \alpha_j \hat{k}_2(\boldsymbol{x}_i,\boldsymbol{x}_j) \quad (6)$$

并满足

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \quad (7)$$

$$0 \leqslant \alpha_i \leqslant c \qquad i = 1,2,\cdots,n$$

解此拉格朗日方程,即可得到以下的输出形式:

$$y(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_i)\rangle + b\right) \quad (8)$$

因为 $\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_i)\rangle$ 的取值与 $\boldsymbol{x}$ 所属的类有关,如果输入向量 $\boldsymbol{x}$ 的类属已知,则可以通过直接带入求解来验证分类的对错;如果 $\boldsymbol{x}$ 所属的类未知,则式(8)就不能直接通过带入 $\boldsymbol{x}$ 来求解类别,为此本文利用如下的定理 2 来解决该问题。

定理 2　对于未知类属的输入向量,基于变换核(式(1))的支持向量机分类算法(式(8))与式(9)等价,即

$$y(\boldsymbol{x}) = \begin{cases} \lambda_1\sum_{i=1}^{n}\alpha_i y_i k_N + b + \frac{1}{2}\sum_{i=1}^{n}\alpha_i(1-\lambda_1) & y > 0 \\ \lambda_1\sum_{i=1}^{n}\alpha_i y_i k_N + b - \frac{1}{2}\sum_{i=1}^{n}\alpha_i(1-\lambda_1) & y < 0 \\ \lambda_1\sum_{i=1}^{n}\alpha_i y_i k_N + b & y = 0 \end{cases} \quad (9)$$

证明　不失一般性,设 n 个训练向量分为两类,$\boldsymbol{x}_i^{(1)}$ 对应大于 0 的类,$i = 1,2,\cdots,l$;$\boldsymbol{x}_j^{(-1)}$ 对应小于 0 的类,$j = 1,2,\cdots,m$,并且 $l + m = n$。

假设 $y(\boldsymbol{x}) > 0$,则

$$\begin{aligned} y(\boldsymbol{x}) &= \left(\sum_{i=1}^{n}\alpha_i y_i\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x})\rangle + b\right) \\ &= \left(\sum_{i=1}^{l}\alpha_i y_i k_1 + \sum_{j=1}^{m}\alpha_j y_j k_2 + b\right) \\ &= \left(\sum_{i=1}^{l}\alpha_i y_i(\lambda_1 k_N + 1 - \lambda_1) + \sum_{j=1}^{m}\alpha_j y_j \lambda_1 k_N + b\right) \\ &= \left(\sum_{i=1}^{n}\alpha_i y_i(\lambda_1 k_N) + \sum_{i=1}^{l}\alpha_i(1-\lambda_1) + b\right) \end{aligned}$$

因为 $0 \leqslant k_N \leqslant 1$,又由式(7)得

$$\begin{aligned} y(\boldsymbol{x}) &= \left(\sum_{i=1}^{n}\alpha_i y_i \lambda_1 k_N + \sum_{i=1}^{l}\alpha_i(1-\lambda_1) + b\right) \\ &= \left(\lambda_1\sum_{i=1}^{n}\alpha_i y_i k_N + \frac{1-\lambda_1}{2}\sum_{i=1}^{n}\alpha_i + b\right) \end{aligned}$$

假设 $y(\boldsymbol{x}) < 0$ 时,则

$$\begin{aligned} y(\boldsymbol{x}) &= \left(\sum_{i=1}^{n}\alpha_i y_i\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x})\rangle + b\right) \\ &= \left(\sum_{i=1}^{l}\alpha_i y_i k_2 + \sum_{j=1}^{m}\alpha_j y_j k_1 + b\right) \\ &= \left(\sum_{j=1}^{m}\alpha_j y_j(\lambda_1 k_N + 1 - \lambda_1) + \sum_{i=1}^{l}\alpha_i y_i \lambda_1 k_N + b\right) \\ &= \left(\sum_{i=1}^{n}\alpha_i y_i(\lambda_1 k_N) - \sum_{j=1}^{m}\alpha_j(1-\lambda_1) + b\right) \\ &= \left(\lambda_1\sum_{i=1}^{n}\alpha_i y_i k_N - \frac{1-\lambda_1}{2}\sum_{i=1}^{n}\alpha_i + b\right) \end{aligned}$$

当 $y(\boldsymbol{x}) = 0$ 时,则由核变换式(1)可知,因为 y 值与 α_i 非零的已有训练数据不属同一类,所以它的核是 $\lambda_1 k_N$,而式(8)则变为

$$\begin{aligned} y(\boldsymbol{x}) &= \left(\sum_{i=1}^{n}\alpha_i y_i\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x})\rangle + b\right) \\ &= \left(\sum_{i=1}^{n}\alpha_i y_i \lambda_1 k_N + b\right) \end{aligned}$$

因此,式(8)等价为下式,即与式(9)等价:

$$y(\boldsymbol{x}) = \begin{cases} \lambda_1\sum_{i=1}^{n}\alpha_i y_i k_N + b + \frac{1}{2}\sum_{i=1}^{n}\alpha_i(1-\lambda_1) & y > 0 \\ \lambda_1\sum_{i=1}^{n}\alpha_i y_i k_N + b - \frac{1}{2}\sum_{i=1}^{n}\alpha_i(1-\lambda_1) & y < 0 \\ \lambda_1\sum_{i=1}^{n}\alpha_i y_i k_N + b & y = 0 \end{cases}$$

4　实验及分析

为验证本文算法的效果,本文首先考虑 2 维向量的人工数据集:$\{\boldsymbol{x}_i, i = 1,2,\cdots\}$;$\boldsymbol{x}_i = (x_i, y_i)$,在空间 $[x_1,x_2] \times [y_1,y_2]$ 上均匀分布,在此空间上,两类数据被一个非线性边界 $y - \frac{y_1 + y_2}{2} = \sin\left(\pi\left(x - \frac{x_1 + x_2}{2}\right)\right)$ 分割。这样就可利用支持向量机分类器产生一个新的非线性边界 $y = \mu(x)$。

在实验中,可随机均匀产生 120 个数据点,并计算出它们的所属类别。试验时,首先对数据进行归一化处理;即数据的输入空间归一化到 $[-1,+1] \times [-1,+1]$ 区间上;然后采用高斯核函数,利用区域搜索法寻找高斯核函数的最优值。对于训练数据首先加入 10% 的高斯噪声,并进行分类;然后采用搜索法找寻高斯核的最优参数。3 种分类结果如表 1 所示。从表 1 可以看出,基于核变换的方法支持向量最少,不仅无分类误差,且泛化能力最强。

随后,本文采用 Wisconsin breast cancer 数据[13]进行了测试。每个输入数据包含 9 维的医学特性输入空间,数据集合包含 699 个样例,其中包括丢失的部分信息。测试时,首先随机选取 200 个训练样本,其中 50 个实验样本用于优化参数,150 个样本用于模型测试;然后利用 200 个训练数据和 50 个测试样本来得到优化的高斯参数 $p_1 = 0.4$;最后利用传统的支持向量机算法和核变换的方法来得到最优分类超平面的各个参数值(如表 2 所示)。

表 1 3 种分类结果与参数比较(10%高斯噪声)

Tab. 1 Results with 3 different methods (10% Gaussian noise)

分类方法	支持向量个数	核参数	分类误差	泛化能力	λ_1
支持向量最少的方法	26	0.75	4	0.07	1
无分类误差的方法	39	0.25	1	0.12	1
基于变换核的方法	15	0.75	0	0.66	0.78

表 2 实际数据的分类结果比较

Tab. 2 Comparison results with real data sets

分类方法	核参数	支持向量个数	分类误差	泛化能力	λ_1
基于变换核的方法	0.60	91	6	0.86	0.75
传统支持向量机的方法	0.40	109	7	0.22	1

从表 2 可以看出,虽基于核变换的方法的支持向量最少,分类误差最小,泛化能力最强,但效果不如有误差和噪声时的仿真实验结果显著,其主要原因是该实验数据是 9 维输入向量,训练和测试数据都相对较少。由于数据相对稀疏,所以用核变换的方法的分类效果不是特别显著,但是仍可以看出分类效果还是有所提高。

5 结 论

本文提出了一种简单有效的基于核变换的支持向量机分类算法。该算法首先根据输入数据的类属性,对核函数进行平移变换,使异类数据的距离增大,即相似性减小;然后使同类数据的距离减小,即相似性增大,从而创造出一个新的核矩阵;最后利用变换后的核来对求解分类数据特征空间得到的超平面方程进行分类。仿真实验和真实的医学数据实验都证明了基于变换核的分类方法不仅可以提高系统的分类性能,而且可增强系统的泛化能力、降低噪声的干扰和增强分类结果的鲁棒性。

参考文献(References)

1 Amari S, Wu S. Improving support vector machine classifiers by modifying kernel functions [J]. Neural Networks, 1999, **12**(6): 783 ~ 789.

2 Kwok James T, Tsang Ivor W. Learning with idealized kernels[A]. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)[C], Washington, DC, USA, 2003:1233 ~ 1237.

3 Xing E, Ng A, Jordan M, *et al*. Distance metric learning, with application to clustering with side-information[A]. In: Advances in Neural Information Processing System[C], Cambridge, MA, USA, 2002:1 ~ 9.

4 Graf Arnulf B A, Smola Alexander J, Borer S. Classification in a normalized feature space using support vector machines [J]. IEEE Transactions on Neural Networks, 2003, **14**(3):597 ~ 605.

5 Lin Y, Lee Y, Wahba G. Support vector machines for classification in nonstandard situations [J]. Machine Learning, 2002, **46**(3):191 ~ 202.

6 Chiang J H, Hao P Y. A new kernel based fuzzy clustering approach: support vector clustering with cell growing [J]. IEEE Transactions on Fuzzy Systems, 2003, **11**(4):518 ~ 527.

7 Zhang Xiang, Xiao Xiao-ling, Xu Guang-you. Determination and analysis of fuzzy membership for SVM[J]. Journal of Image and Graphics, 2006, **11**(8):1188 ~ 1192. [张翔,肖小玲,徐光佑. 模糊支持向量机中隶属度的确定与分析 [J]. 中国图象图形学报, 2006,**11**(8):1188 ~ 1192.]

8 Chen Y, Wang J Z. Support vector learning for fuzzy rule based classifications system [J]. IEEE Transactions on Fuzzy Systems, 2003, **11**(6):716 ~ 728.

9 Cristianini N, Shawe-Tayor J, Elisseeff A, *et al*. On kernel target alignment [A]. In: Advances in Neural Information Processing System [C], Cambridge, MA, USA, 2002:367 ~ 373.

10 Herbrich R, Graepel T. A PAC-bayesian margin bound for linear classifiers: why SVM's work [J]. IEEE Transactions on Information Theory, 2002:**48**(12): 3140 ~ 3150.

11 Vapnik V, Chapelle O. Bounds on error expectation for support vector machines [J]. Neural Computation, 2000, **12**(9):585 ~ 592.

12 Scholkopf B, Smola A. Learning with kernels: support vector machines, regularization, optimization, and beyond [M]. Cambridge, MA,USA: MIT Press, 2002: 12 ~ 45.

13 Wisconsin breast cancer [EB/OL]. http://www.ics.uci.edu/~mlearn/ MLRepository.html.