基于密度法的双隶属度模糊支持向量机

李 雷,周蒙蒙,鲁延玲
(南京邮电大学 理学院,江苏 南京 210003)

摘 要:针对现实环境中样本集越来越大,并且往往含有大量噪声和野值,导致传统模糊支持向量机的训练时间和分类识别率降低的问题,提出基于密度法的双隶属度模糊支持向量机,即靠近类中心的样本点隶属度由其到类中心的距离确定,远离类中心的样本点隶属度由其邻域内同类异类样本点数量的比例确定。从理论和实证两个方面分析文中方法与以往基于密度的模糊支持向量机(DFSVM)相比,该方法不但降低了算法的复杂度,并且提高了支持向量机的分类精度。

关键词:模糊支持向量机;双隶属度;密度;类中心
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2009)12-0044-03

Fuzzy Support Vector Machine Based on Density with Dual Membership

LI Lei, ZHOU Meng-meng, LU Yan-ling
(School of Sciences, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In this paper, an improved fuzzy membership function determination is proposed to train the fuzzy support vector machine (FSVM) for classification which the sample set in reality environment is increasing, and it often contains a lot of noise and outliers. In the improved algorithm, the sample points have the different types of memberships in different regions. That is, the membership of the sample point near by the class centers is determined by the distance between the point and its class center, and the membership of the sample point far away the class centers is determined by the proportion between the number of its congeneric points and the number of its heterogeneous points in its neighborhood. The dual membership is introduced to reduce the algorithm complexity and shorten its training time compared with fuzzy support vector machine based on density, at the same time the algorithm well improves the SVM's accuracy rate.

Key words: fuzzy support vector machine; dual membership; density; class center

0 引 言

支持向量机(Support Vector Machine,SVM)[1]基于统计学理论的VC维(Vapnik Chervonenks dimension)理论和结构风险最小化原理(Structural Risk Minimization),很大程度上克服了传统机器学习中的非线性、维数灾难及局部极小等问题。目前,在处理小样本问题时,SVM的泛化能力最好,它能较好地解决小样本、非线性、高维数和局部极小点等实际存在的问题;又因为它采用了核函数思想,把非线性空间的问题转换到线性空间,这样就大大降低了算法的复杂度。由于SVM出色的学习性能,该技术已成为学术研究的热点,并在许多领域中得到了应用[2~6]。但是,目前SVM还存在很多的局限性。比如,如何将解决海量数据的分类问题,如何克服对孤立点和噪音数据的敏感问题[7]。

为了解决支持向量机在实际应用中遇到的问题,引入了模糊支持向量机(FSVM)[8]。在模糊支持向量机中,隶属度函数的设计是整个算法的关键,这要求隶属度函数必须能够客观、准确地反映系统样本点的不确定性;具有良好的去噪声和去野值能力;同时为了提高计算机的运行效率,必须严格控制算法的隶属度函数的复杂度。在文献[9]中提出了基于类中心的隶属度设计思想,各样本点的隶属度根据其到类中心的距离确定,但其中类半径的确定对野值和噪声非常敏感。文献[10]提出了基于密度法的隶属度设计思想,各样本点的隶属度由其邻域内同类异类样本点的比例确定,但这种方法也有其缺陷。首先,在类中心区域,样本点的隶属度全部接近于1,不能准确地反映不同样本点对支持向量机的不同影响;另外,每个样本点都要在其邻域内寻找同类异类样本点,算法复杂度大高,计算机运行时间明显增加。

收稿日期:2009-03-10;**修回日期**:2009-06-18
基金项目:国家自然科学基金项目(10371106, 10471114);江苏省高校自然科学基金项目(04KJB110097, 08KJB520023);南京邮电大学攀登计划项目(NY207064)
作者简介:李 雷(1958-),男,安徽砀山人,教授,研究方向为智能信号处理、非线性分析与计算智能。

在文中提出了一种新的模糊隶属度的确定方法,即靠近类中心的样本点隶属度由其到类中心的距离确定,远离类中心的样本点隶属度由其邻域内同类异类样本点的比例确定;同时为了降低噪声和野值对类半径的影响,对类半径的确定做了改进。实验表明,基于密度法的双隶属度模糊支持向量机较其它算法提高了分类识别率,降低了计算机运行时间。

1　模糊支持向量机(FSVM)

在传统的 SVM 理论中,训练过程对于那些远离它们所属类的训练点是十分敏感的,但在 SVM 的训练过程中,所有的训练点都被平等对待。这样就导致了 SVM 对某些特殊情形的点过分敏感,这种情形即是所谓的过学习现象。在很多实际应用问题中,不同训练点对分类结果的影响是不同的。一般来说,训练集中存在某些点对分类结果的影响很大,同时也存在一些点对分类结果的影响很小,甚至是微不足道的。因此,在处理分类问题时,必须将那些重要的点正确分类,并且可以忽略那些微不足道的点。

模糊支持向量机的核心思想是引入模糊隶属度,根据不同输入样本对分类的贡献不同,赋以相应的隶属度,这样可减小野值和噪声的影响,提高 SVM 的分类性能。在此思想下,训练点不再严格属于两类中的某一类。而有可能存在下列情况,某一个训练样本 80% 的可能属于某一类,20% 的可能不属于这一类。另外一个样本也可能 70% 的属于某一类,30% 的可能不属于这一类。也就是,对于每个样本点 x_i,存在一个与之对应的模糊成员函数 μ_i,$0 < \mu_i \leqslant 1$,把 μ_i 称为隶属度函数。

给定训练样本集 $T = \{x_1, y_1, \mu_1\}, \{x_2, y_2, \mu_2\}, \cdots, \{x_l, y_l, \mu_l\}$,其中 $x_i \in R^N$ 为样本特征,$y_i \in \{-1, 1\}$ 为类标识,$\mu_i \in (0,1]$ 表示 x_i 属于 y_i 的程度,称之为样本点 $\{x_l, y_l, \mu_l\}$ 的隶属度,$i = 1,2,\cdots,l$,从而利用模糊支持向量机求解最有超平面的优化问题为:

$$\min \frac{1}{2} \| \omega \|^2 + C\sum_{i=i}^{l} \mu_i \xi_i$$

s.t.

$$\begin{cases} y_i \cdot ((\omega \cdot \varphi(x_i)) + b) - 1 + \xi_i \geqslant 0 \\ \xi_i \geqslant 0 \end{cases} \quad i = 1,2,\cdots,l$$

引入 Lagrange 乘子:

$$L(\omega, b, \xi_i) = \frac{1}{2} \| \omega \|^2 + C\sum_{i=1}^{l} \mu_i \xi_i - \sum_{i=1}^{l} \alpha_i [y_i \cdot (\omega \cdot \varphi(x_i) + b) - 1 + \xi_i] - \sum_{i=1}^{l} \beta_i \xi_i$$

根据 KKT 条件有:

$$\frac{\partial L(\omega, b, \xi_i)}{\partial \omega} = \omega - \sum_{i=1}^{l} \alpha_i y_i \varphi(x_i) = 0$$

$$\frac{\partial L(\omega, b, \xi_i)}{\partial b} = -\sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\frac{\partial L(\omega, b, \xi_i)}{\partial \xi_i} = C\mu_i - \alpha_i - \beta_i = 0$$

由此求解问题变为下面的一个二次优化问题:

$$\max \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

s.t.

$$\begin{cases} 0 \leqslant \alpha_i \leqslant \mu_i C \\ \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases} \quad i = 1,2,\cdots,l$$

其中,$C > 0$ 为惩罚参数,表示对错分样本惩罚的程度;μ_i 为样本点的模糊隶属度,$\xi_i \geqslant 0$ 为松弛变量,$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$ 为满足 Mecer 核定理的核函数。

这样得到其对应的最优分类面的决策函数为:

$$f(x) = \text{sign}[\sum_{i=1}^{l} \alpha_i y_i K(x, x_i) + b]$$

2　模糊隶属度函数算法

给定训练样本集 $T = \{x_1, y_1, \mu_1\}, \{x_2, y_2, \mu_2\}, \cdots, \{x_l, y_l, \mu_l\}$,式中各符号如前节所述。

2.1　定　义

样本点之间的距离:

$D(x_i, x_j) = \| x_i - x_j \|$

样本点的同类点密度、异类点密度分别为:

$\rho^+(x_i, R) = |\{x_j \mid D(x_j, x_i) \leqslant R, y_j = y_i\}|$

$\rho^-(x_i, R) = |\{x_j \mid D(x_j, x_i) \leqslant R, y_j \neq y_i\}|$

式中 $|E|$ 表示集合 E 的势,即集合 E 中元素的个数,R 为可调节的样本点邻域半径。

设正负类样本点的类中心分别为:O^+, O^-。定义类中心为:

$$O^+ = \frac{\sum_{y_i=1} x_i}{l^+} \qquad O^- = \frac{\sum_{y_i=-1} x_i}{l^-}$$

式中 l^+、l^- 分别为正负样本点的个数。

调节 O^+、O^- 的邻域半径,使得

$\rho(O^+) = |\{x_j \mid D(x_j, O^+) \leqslant R^+, y_j = 1\}| = a\% l^+$

$\rho(O^-) = |\{x_j \mid D(x_j, O^-) \leqslant R^-, y_j = -1\}| = a\% l^-$

$0 < a < 100$,为可调参数。

得到正负类样本的类半径分别为 R^+、R^-。这样,由正(负)类中心和正(负)类半径组成的超球只覆盖了 $a\%$ 的正(负)类样本点,大大降低了噪声和野值对正(负)类半径的影响,如图 1 所示。

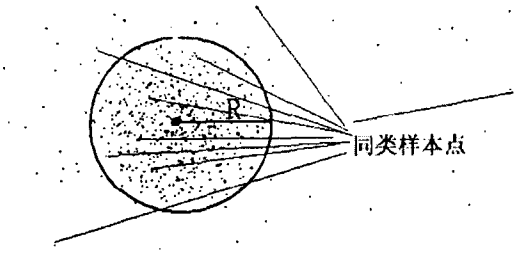

图 1 类半径的确定

2.2 隶属度函数的确定

根据到其类中心的距离把样本点分为两类。两类样本点采用两种不同的隶属度设计方法。

2.2.1 靠近类中心的样本点的隶属度

靠近类中心的样本点是指符合下式的点：

$$D(x_i,O^+)\leqslant bR^+,y_i=+1$$

or

$$D(x_i,O^-)\leqslant bR^-,y_i=-1$$

$0<b\leqslant 1$ 为可调节参数。

隶属度函数为：

$$\mu_i=\begin{cases}\dfrac{1}{1+\dfrac{cD(x_i-O^+)}{R^+}} & y_i=+1\\ \dfrac{1}{1+\dfrac{cD(x_i-O^-)}{R^-}} & y_i=-1\end{cases}$$

$c>0$ 为可调节参数。

2.2.2 远离类中心的样本点的隶属度

远离类中心的样本点是指符合下式的点：

$$D(x_i,O^+)>bR^+,y_i=+1$$

or

$$D(x_i,O^-)>bR^-,y_i=-1$$

隶属度函数为：

$$\mu_i=\frac{d\rho^+(x_i,R)}{\rho^+(x_i,R)+\rho^-(x_i,R)}$$

$0<d\leqslant 1,R>0$ 为可调节参数。

这样该算法既准确地反映了不同样本点对支持向量的影响，又降低了算法的复杂度。

3 数值实验及结论

实验用三组数据均从 UCI 数据库下载，其中每组实验用训练样本 1000 个，测试样本 1800 个；编程语言采用 MATLAB；实现硬件环境为 TK－53，1.73G×2 CPU，1G 内存。所使用的距离定义均为欧氏距离；核函数采用径向基函数；文中算法所涉及到的各参数根据经验及反复实验多次修正，最后选用较优结果，且三组实验所用的参数值也不相同。文中算法与普通支持向量机、文献[10]算法的实验结果如表 1 所示。

从表 1 可以看出，文中所提出的基于密度法的双重隶属度模糊支持向量机(DDFSVM)较 DFSVM 和 FSVM 的分类精度有了提高，同时较 DFSVM 的运行时间大大降低。实验表明文中算法较其它算法减少了噪声对分类的影响，同时也降低了算法复杂度，缩短了运行时间，提高了模糊支持向量机的使用范围。

表 1 三种算法分别对三种数据的实验结果对比

	SVM		DFSVM		DDSVM	
数据集	分类精度	运行时间	分类精度	运行时间	分类精度	运行时间
Australian	72.56%	0.42s	79.78%	7.35s	92.42%	1.64s
Diabetics	69.03%	1.15s	80.34%	9.01s	93.48%	3.63s
German	79.07%	1.49s	83.72%	11.42s	96.20%	3.91s

4 结束语

提出了基于密度法的双隶属度模糊支持向量机，实验结果表明，文中提出的算法较其它支持向量机算法明显提高了分类性能。怎样确定区分靠近类中心的点与远离类中心的点，即如何选取较优的参数 b；以及如何界定样本集中的噪声野值，即如何选取较优的参数 a 都是作者未来的研究方向。

参考文献：

[1] 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.

[2] Chen Jiaming, Li Lei, Nie Lingye. Wavelet image compression by using hybrid kernel SVM [C] // Proceeding of ICMLC2008. Kunming: [s. n.], 2008: 3056－3060.

[3] Cui Jianggui, Li Zhonghai. The Application of Support Vector Machine in Pattern Recognition[C]//2007 IEEE International Conference on Control and Automation. Guangzhou: [s. n.], 2007: 3135－3138.

[4] 王玉震，李 雷. 基于 SVR 的图像增强方法[J]. 计算机技术与发展，2009，19(1)：60－62.

[5] 陈 潇，李 雷，范小岗. 基于支持向量机的非线性多用户检测[J]. 西安邮电学院学报，2008，13(1)：37－40.

[6] 张成伟，郑 诚. 基于改进 VSM 的文本信息检索研究[J]. 计算机技术与发展，2009，19(1)：71－73.

[7] Lin Chunfu, Wang Shengde. Fuzzy support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464－471.

[8] Song Qing, Hu Wenjie, Xie Wenfang. Robust support vector machine with bullet hole image classification[J]. IEEE transactions on systems, man and cybernetics－part C: applications and reviews, 2002, 32(4): 440－448.

[9] Zhang Xuegong. Using class－center vectors to build support vector machines[C] // Neural Networks for Signal Processing IX－Proc of the 1999 IEEE Workshop. Wisconsin: IEEE Inc, 1999: 33－37.

[10] 安金龙，王正欧，马振平. 基于密度法的模糊支持向量机[J]. 天津大学学报，2004，37(6)：544－548.