# 一种改进的模糊支持向量机算法

刘三阳,杜 喆
(西安电子科技大学 理学院,陕西 西安 710071)

**摘 要**:模糊隶属度函数设计是模糊支持向量机中的关键步骤. Lin & Wang 提出的基于类中心距离的模糊隶属度设计方法,不能从样本集中有效区分噪声或野值点,而且可能降低支持向量的隶属度. 针对上述不足,提出一种改进的隶属度函数设计方法. 通过引入一个半径控制因子,充分利用样本间的信息,更加合理地设计样本的模糊隶属度. 与基于类中心的隶属度方法相比,该方法在不增加时间复杂度的情况下,通过数值实验表明了方法的优势,大大提高了模糊支持向量机的分类精度.

**关键词**:模糊支持向量机;隶属度函数;分类
**中图分类号**: TP181 **文献标识码**: A **文章编号**: 1673-4785(2007)03-0030-04

# An improved fuzzy support vector machine method

LIU San-yang, DU Zhe
(School of Science, Xidian University, Xi'an 710071, China)

**Abstract**: A design that improves the classifying ability of an SVM by improving the assignment of fuzzy membership is an important step in solving the fuzzy SVM problem. In this paper, a radius controlling factor is introduced to assign sample membership more accurately. This technique can distinguish noise and outliers and increase accuracy of membership in support vectors, compensating for the disadvantages of assigning fuzzy membership based on the distance between a sample and its cluster center proposed by Lin and Wang. Experimental results verify the effectiveness of this method.

**Key words**: fuzzy support vector machine; membership function; classification

支持向量机(support vector machines, SVMs)是一种基于统计学习理论的新的机器学习方法[1-2],近来,由于其表现出了很强的泛化能力,能很好地克服维数灾难和过学习等传统算法所不可回避的问题,而日益受到广泛重视. 支持向量机的处理机制决定了其对训练样本内的噪音和孤立点特别敏感. 针对这种情况,Lin 等学者提出了模糊支持向量机方法(FSVM)[3-5],将模糊技术应用于支持向量机中,对不同的样本采用不同的惩罚权系数,使得在构造目标函数时,不同的样本有不同的贡献,对含有噪声或野值的样本赋予较小的权值,从而达到消除噪声与野值样本影响的目的.

在采用模糊技术处理时,隶属度函数的设计是整个模糊算法的关键,这要求隶属度函数必须能够客观、准确地反映系统中样本存在的不确定性. 目前,构造隶属度函数的方法很多[6-7],但还没有一个可遵循的一般性准则. 在对实际情况进行处理时,通常需要针对具体问题,根据经验来确定合理的隶属度函数. 关于隶属度函数,不少学者在这方面作了一些研究,但主要是基于样本到类中心之间的距离来度量其隶属度的大小[3,6]. 然而,在依据样本到类中心之间距离的角度确定样本的隶属度时,有时并不能将含噪声或野值样本从有效样本集中区分出来,以致将含噪声或野值样本与有效样本赋予相同的隶属度. 其主要原因是在依据样本到类中心之间距离的角度确定样本的隶属度时,没有考虑样本之间的关系,而仅仅考虑样本与类中心之间的距离.

## 1 模糊支持向量机(FSVM)

设输入的样本集合$\{x_i\} \in R^n$ 由 2 类点组成,标号 $y_i = 1$ 或 - 1,则训练样本集为$(x_i, y_i)$, $i = 1, 2, 3,$

收稿日期:2006-10-08.
基金项目:国家自然科学基金资助项目(60574075).

…, $n$. 将线性可分与线性不可分情况统一考虑,引入松弛因子 $\xi_i$,如果分类超平面为 $\omega \cdot x_i + b = 0$,满足约束:

$$y_i[\omega \cdot x_i + b] - 1 + \xi_i \geq 0.\ i = 1,2,3,\cdots,n. \tag{1}$$

当 $\{\xi_i\} = 0$,表示线性可分,否则为线性不可分情况.

如果考虑非线性情况,假设存在影射 $\phi$ 将样本 $x$ 从低维空间影射到高维空间 $H$ 中的 $\phi(x)$,则在高维空间中,训练样本集合则变为 $(\phi(x_i), y_i)$,在这里可以引人一个模糊因子 $s_i$, $0 < s_i \leq 1$, $i = 1,2,3,\cdots,n$. $s_i$ 表示第 $i$ 个样本属于正常的程度. 此时,原先的训练集合就可以转化为带有模糊因子的训练样本集合 $(\phi(x_i), y_i, s_i)$, $s\xi_i$ 则为带有不同权重的松弛因子. 因此,求解支持向量机最优超平面问题就可以转化为下面的优化问题:

$$\begin{aligned} &\min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} s_i\xi_i, \\ &\text{s.t.}\quad y_i[\omega \cdot \phi(x_i) + b] - 1 + \xi_i \geq 0, \\ &\qquad \xi_i \geq 0 \quad i = 1,2,3,\cdots,n. \end{aligned} \tag{2}$$

式中:$C$ 为常数. $s_i$ 越小,则相应的样本 $x_i$ 在对式(2)优化问题所起的作用就越小.

求解上面的优化问题,可以先构造下面的拉格朗日函数,即

$$L(\omega,b,\alpha,\beta) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} s_i\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\omega \cdot \phi(x_i) + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta\xi_i. \tag{3}$$

要求出上述函数的鞍点,则需满足下面的条件:

$$\frac{\partial L(\omega,b,\alpha,\beta)}{\partial \omega} = \omega - \sum_{i=1}^{n}\alpha_i y_i \phi(x_i) = 0. \tag{4}$$

$$\frac{\partial L(\omega,b,\alpha,\beta)}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0. \tag{5}$$

$$\frac{\partial L(\omega,b,\alpha,\beta)}{\partial \xi_i} = s_i C - \alpha_i - \beta_i = 0. \tag{6}$$

将式(4)、(5)、(6)代入式(3),则求解优化问题变成求解下面的二次规划问题:

$$\max W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j). \tag{7}$$

满足的约束条件为

$$\sum_{i=1}^{n}\alpha_i y_i = 0, 0 \leq \alpha_i \leq s_i C, i = 1,2,3,\cdots,n. \tag{8}$$

式中:$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ 为核函数. Karush-Kuhn-Tucker 条件为

$$\begin{aligned} &\alpha_i[y_i(\omega \cdot \phi(x_i) + b) - 1 + \xi_i] = 0, \\ &(s_i C - \alpha_i)\xi_i = 0, i = 1,2,3,\cdots,n. \end{aligned} \tag{9}$$

可以得到 $\alpha_i = 0$ 对应的训练样本点 $x_i$ 是被完全正确分类的,$\alpha_i > 0$ 对应的 $x_i$ 被称为支持向量(support vector). $0 < \alpha_i < s_i C$ 对应的训练样本 $x_i$ 在超平面的间隔之间. 而 $\alpha_i = s_i C$ 所对应的样本 $x_i$ 是被错分的. 反之也是成立的.

从上面的公式可以看出,$s_i C$ 表示样本在 $x_i$ 训练支持向量机时的重要程度:$s_i C$ 越大,表示样本 $x_i$ 被错分的可能性越小,分类超平面与各类样本间的距离越小;反之,样本 $x_i$ 被错分的可能性越大,分类超平面与各类样本间的距离也越大. 对于孤立点或噪音样本,如果能够使其对应的 $s_i$ 很小,从而 $s_i C$ 很小,则此样本对支持向量机的训练作用就大为减小了,其结果便是大大降低了他们对训练支持向量机的影响. 可见模糊因子 $s_i$ 的确定成为决定这种模糊支持向量机工作性能好坏的关键问题.

## 2　改进的隶属度函数算法

为了减小孤立点数据对支持向量机分类的影响,文献[3]引入的模糊因子,使样本对支持向量机分类所起的作用随着训练样本集合中样本远离类别的几何中心点而逐渐减小,而最优超平面主要是由距最优超平面距离最近的点即支持向量(包含于 2 类样本相对边界向量)来确定. 由于这些支持向量都位于 2 类样本的相对边界上,距 2 类类中心点的距离都较远,如果按照文献[3]提出的减小孤立点作用的方法,在减小孤立点作用的同时,也大大减小了支持向量对分类超平面的作用,其最终结果将会使所获得的分类超平面偏离最优分类超平面,从而影响了支持向量机的分类性能.

实际上,文献[3]中构造了一个以类中心为球心,样本距中心最大距离为半径的超球,将所有的同类样本包含在内. 本文将文献[3]中的方法进行改进,提出一种改进的模糊支持向量机(ratio fuzzy SVM). 因为对于 2 类问题,分类平面一定处于 2 个类中心之间,而引进一个控制因子,来控制以类中心为球心的球半径大小,可将半径之外的点赋予一个很小的隶属度,这样可以有效区分噪声和野值点. 再将样本与球心的距离与半径的比定义为隶属度,这样通过调节半径控制因子,使支持向量大体处于球面附近,从而提高支持向量的隶属度.

与文献[3]类似,使用正类样本的均值作为正类的中心,记为 $x_+$,负类样本的均值作为负类的中心,记为 $x_-$. 定义正类的半径:$R_+ = \max_{\{x_i : y_i = +1\}} |x_i - x_+|$,负类的半径:$R_- = \max_{\{x_i : y_i = -1\}} |x_i - x_-|$. 2类中心的距离为 $T = |x_+ - x_-|$. 每个正类样本到正类中心的距离为 $D_i^+ = |x_+ - x_i|$,每个负类样本到负类中心的距离为 $D_i^- = |x_- - x_i|$. $\sigma$为一个事先给定的很小的正数,作为噪声和孤立点隶属度. $\varepsilon$为引入的半径控制因子,满足 $\varepsilon > 0$,使 $T \cdot \varepsilon < R_+$ 和 $T \cdot \varepsilon < R_-$. 则隶属度函数定义为

$$s_i^+ = \begin{cases} \dfrac{\delta + D_i^+}{R_+}, & D_i^+ \leq T \cdot \varepsilon, \\ \sigma, & D_i^+ > T \cdot \varepsilon, \end{cases}$$

$$s_i^- = \begin{cases} \dfrac{\delta + D_{+i}}{R_-}, & D_i^- \leq T \cdot \varepsilon, \\ \sigma, & D_{+i} > T \cdot \varepsilon. \end{cases}$$

式中:$\delta$是一很小的正数,为了保证 $s_i > 0$.

## 3 数值实验及结果分析

文中采用 Matlab 语言编写的一个 SVM 工具箱,在 CPU2.6G,内存为 512M 的微机上进行数值实验. 比较指标是对测试集的分类精度(%)和训练时间(s).

首先对人工数据进行实验. 训练样本和测试样本都是随机产生2类二维含噪声的样本点,其中正类数目为 200 和 100,负类数目也是 200 和 100. 分别采用标准 SVM 算法,文献[3]中的 FSVM 算法,以及文中提出的 RFSVM(ratio fuzzy SVM)算法,进行数值实验. 在相同的参数条件下,分别运行 10 次,取平均值,比较见表 1. 结果表明本文的算法有很大的优越性.

表1 SVM、FSVM 和 RFSVM 在人工数据集上的性能比较

Table 1 Performance comparison of SVM、FSVM and RFSVM on artificial data set

| | SVM | FSVM | RFSVM |
|---|---|---|---|
| 错误数(率)/% | 11.2 (5.6) | 10.2 (5.1) | 8.7 (4.4) |
| 运行时间/s | 1.20 | 1.22 | 1.22 |

然后对 UCI 数据库中的真实数据集进行实验,也表现出良好的分类效果. 比较结果见表 2. 从 Breast-cancer 数据实例也可看出文献[3]中的设计方法未必很好.

需要指出的是:RFSVM 与 SVM 相比因需要额外计算隶属度函数,执行时间相对多一点. 这和 FSVM 类似,但可得到更高的分类精度作为回报. 与 FSVM 相比,算法中仅加入一个判断语句,时间复杂度上并没有增加,运行时间基本相同,而且得到更高的分类精度,这也正是 RFSVM 优势所在.

表2 SVM、FSVM 和 RFSVM 在真实数据集上的性能比较

Table 2 Performance comparison of SVM, FSVM and RFSVM on real data set

| 数据集(训练/测试) | | SVM | FSVM | RFSVM |
|---|---|---|---|---|
| Breast (200/700) | 错误率/% | 26.06 | 24.12 | 24.03 |
| | 时间/s | 0.80 | 1.51 | 3.43 |
| Diabetics (468/300) | 错误率/% | 27.02 | 24.01 | 23.57 |
| | 时间/s | 0.82 | 1.54 | 3.50 |
| German (700/300) | 错误率/% | 25.47 | 23.16 | 23.21 |
| | 时间/s | 0.81 | 1.54 | 3.51 |

## 4 结束语

文中针对基于类中心距离的模糊隶属度的设计方法的不足之处,通过引入一个半径控制因子,来控制以类中心为球心的球半径的大小,更加合理地设计样本的模糊隶属度. 一方面大大减弱了孤立点和噪音点对支持向量机最优超分类平面的影响,另一方面,不影响支持向量对最优分类超平面的决定作用. 在抗击孤立点和噪音点的干扰方面,本方法在性能上远远优于文献[3]中的基于类中心点距离的模糊支持向量机方法,取得很好的效果,提高支持向量机分类的泛化能力,而且时间复杂度并没有提高.

## 参考文献:

[1] VAPNIK V. The nature of statistical learning theory [M]. New York: Springer, 1995.

[2] CRISTIANINI N, TAYLOR S J. An introduction to support vector machines [M]. Cambridge: Cambridge University Press, 2000.

[3] LIN C F, WANG S D. Fuzzy support vector machines [J]. IEEE Trans Neural Networks, 2002, 13(2): 464-471.

[4] INOUE T, ABE S. Fuzzy support vector machines for pattern classification [A]. Proceedings of International Joint Conference on Neural Networks [C]. Washington,

D. C. ,2001.
[5]HUANG H P,LIU Y H. . Fuzzy support vector machines for pattern recognition and data mining[J]. International Journal of Fuzzy Systems, 2002,4(3):826 - 835.
[6]ZHANG X G. Using class - center vectors to build support vector machines[A]. Proc IEEE NNSP'99[C]. USA,1999.
[7]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.
AN Jinlong, WANG Zheng'ou, MA Zhenping. Fuzzy support vector machine based on density[J]. Journal of Tianjin University, 2004, 37(6):544 - 548.
[8]边肇祺,张学工. 模式识别(第 2 版)[M]. 北京:清华大学出版社,2000.

**作者简介**:

刘三阳,男,1959 年生,博士,教授,博士生导师,主要研究方向为最优化理论与方法、网络算法及其在通信网中的应用. 先后主持近 20 个科研项目,发表论文 360 多篇.

杜　喆,男,1982 年生,博士研究生,主要研究方向为机器学习与最优化方法.
E-mail:doog2005 @tom. com.

# 中国人工智能学会 2007 年全国学术大会

## 2007 China National Conference on Artificial Intelligence(CAAI - 12)

由中国人工智能学会主办的 2007 年全国学术大会——CAAI - 12(2007 China National Conference on Artificial Intelligence)将于 2007 年 12 月 27 日至 30 日在哈尔滨举行。本届大会由哈尔滨工程大学承办。它将提供中国人工智能界交流最新研究成果的良好舞台。同以往各届大会一样,CAAI - 12 将是我国人工智能界的又一次全国学术盛会。CAAI 2007 的议题涉及人工智能领域各个方面。本次大会将邀请著名科学家做前沿报告。会议期间同时将举办讲座和专题讨论。CAAI 2007 诚请广大人工智能界研究人员投稿。会议的议题主要包括(但不限于此):

1. 人工智能理论基础:脑科学、认知科学、意识学、情感学、泛逻辑学、模糊集、控制论、系统学、协调学、信息 - 知识 - 智能理论、拟人学、离散数学、集对分析与联系数、智能科学哲学;2. 知识科学与知识工程:知识表示、自动推理、专家系统、分布智能;3. 机器学习;4. 智能控制与智能管理;5. 智能机器人与机器人足球;6. 智能信息网络;7. 计算机辅助教育:人工智能教育、人工智能科学普及;8. 机器感知与虚拟现实;9. 生物信息学与人工生命;10. 智能 CAD 与数字艺术;11. 神经网络与计算智能、人工免疫;12. 自然语言处理与理解、机器翻译;13. 可拓工程;14. 粗糙集与软计算;15. 计算机博弈;16. 人工智能应用:智能交通、智能制造、运动控制、模糊设计、智能电力、智能系统工程、智能系统优化;17. 智能产品标准与产业发展。

会议网站:http://caai.cn/caai - 12/　联系电话:0451 - 82518975,13936496483　联系人:张汝波