模糊支持向量机中隶属度的确定与分析

张 翔[1),2)] 肖小玲[3)] 徐光祐[1)]

[1)](清华大学计算机系,北京 100084) [2)](长江大学地球物理与石油资源学院,荆州 434023)

[3)](武汉理工大学计算机科学与技术学院,武汉 430063)

摘 要 针对目前模糊支持向量机方法中,一般使用特征空间中样本与类中心之间的距离关系构建隶属度函数的不足,提出了一种新的有效地反映样本不确定性的隶属度计算方法——基于样本紧密度的隶属度方法。在确定样本的隶属度时,不仅考虑了样本与类中心之间的关系,还考虑了类中各个样本之间的关系,并采用模糊连接度来度量类中各个样本之间的关系。将其应用于模糊支持向量机方法中,较好地将支持向量与含噪声或野值样本区分开。实验结果表明,采用模糊支持向量机方法,其分类错误率比采用支持向量机方法的错误率低,在使用的 3 种隶属度函数中,采用基于紧密度隶属度的模糊支持向量机方法抗噪性能最好,分类性能最强。

关键词 支持向量机 模糊隶属度 紧密度

中图法分类号:TP391.41 **文献标识码**:A **文章编号**:1006-8961(2006)08-1188-05

Determination and Analysis of Fuzzy Membership for SVM

ZHANG Xiang[1),2)], XIAO Xiao-ling[3)], Xu Guang-you[1)]

[1)](*School of Computer Science, Tsinghua University, Beijing* 100084)

[2)](*School of Geophysics and Resources, Yangtze University, Jinzhou* 434023)

[3)](*School of Computer Science and Technology, Wuhan University of Technology, Wuhan* 430063)

Abstract Relative to the fuzzy membership as a function of distance between the point and its class center in feature space for some current fuzzy support vector machines, a new and more effective fuzzy membership as a function of affinity among samples is proposed for the measurement of the inaccuracy of samples. The fuzzy membership is defined by not only the relation between a sample and its cluster center, but also those among samples, which is described by the fuzzy connectedness among samples. The fuzzy membership based on the affinity among samples for support vector machine effectively distinguishes between support vectors and outliers or noises. Experimental results show that the fuzzy support vector machine, based on the affinity among samples is more robust than the traditional support vector machine, and fuzzy support vector machines taken by other two fuzzy memberships.

Keywords support vector machine, fuzzy membership, affinity

1 引 言

尽管支持向量机方法具有较好的推广能力,由于在构造最优分类面时所有的支持向量样本具有相同的作用,当训练样本中含有噪声与野值样本时,这些含有“异常”信息的样本在特征空间中常常位于分类面附近,导致获得的分类面不是真正的最优分类面。另外,在实际应用中,对某些重要的类需要非常高的分类精度,而对其他类分类精度的要求相对低一些,所有这些问题,对于常规的支持向量机方法都无法解决。针对这种情况,提出了模糊支持向量

基金项目:国家自然科学基金项目(60273005);湖北省自然科学基金项目(2004ABA043);中国博士后科学基金(2005038310);湖北省教育厅科学技术研究重点项目(D200612002)

收稿日期:2005-08-25;**改回日期**:2005-10-17

第一作者简介:张翔(1969~),男,副教授,清华大学博士后。主要研究方向为图像处理、模式识别。已在国内外杂志上发表学术论文 20 余篇,其中 SCI 检索 2 篇,EI 检索 10 篇。E-mail:zx_jr_xl@163.com

机方法(FSVM)[1~4],对不同的样本采用不同的惩罚权系数,以致在构造目标函数时不同的样本有不同的贡献,对含有噪声与野值的样本赋予较小的权值,从而在一定的程度上消除了噪声与野值的影响。

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,不同的隶属度函数会对算法的处理结果以及算法实现的难易程度产生不同的影响,这要求隶属度函数必须能客观、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,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。不少的学者在这方面作了一些研究,但在目前的模糊支持向量机方法中,主要是采用基于样本到类中心之间的距离来度量其隶属度的大小[1]。然而,在依据样本到类中心之间距离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。文献[4]在考虑隶属度计算时,对类中有效样本与野值分别进行了考虑,对有效样本采用每个样本到中心点的距离来度量其隶属度的大小,而对野值的隶属度直接赋予一个很小的值。该方法的关键在于需要首先确定野值,而确定野值既非常困难又非常关键[5,6],而错误的野值会严重地影响模糊支持向量机的结果。

本文首先对几种常用的隶属度进行了分析,针对几种隶属度应用中的不足,依据有效样本与含噪声样本或野值在特征空间中的分布特点,研究了一种新的隶属度计算方法——基于紧密度的隶属度函数。在确定隶属度时,不仅考虑了样本与类中心之间的关系,还考虑了类中各个样本之间的关系,有效地将支持向量与含噪声或野值样本区分开,较好地反映了样本的不确定性。

2　隶属度的确定

2.1　常用的几种隶属度函数

2.1.1　基于距离的隶属度函数

一般情况下,确定隶属度大小的基本原则是依据样本所在类中的相对重要性,或对所在类贡献的大小。样本到类中心之间的距离是衡量样本对所在类贡献大小的依据之一。目前,在模糊支持向量机中,基于距离的隶属度函数的确定是将样本的隶属度看作是特征空间中样本与其所在类中心之间距离的函数[1]。

设 $\bar{\boldsymbol{x}}$ 为类中心,r 为类半径,由下式确定

$$r = \max_i \| \boldsymbol{x}_i - \bar{\boldsymbol{x}} \| \tag{1}$$

依据距离确定隶属度时,类中各样本的隶属度为

$$\mu(\boldsymbol{x}_i) = 1 - \frac{\| \boldsymbol{x}_i - \bar{\boldsymbol{x}} \|}{r} + \delta \tag{2}$$

其中,$\delta > 0$ 是预设的一个很小的常数,避免出现 $\mu(\boldsymbol{x}_i) = 0$的情况。

2.1.2　基于 S 型函数的隶属度函数

在基于距离隶属度函数的确定中,将样本的隶属度看作是样本到所在类中心之间距离的线性函数。而实际样本的隶属度与样本到所在类中心的距离之间不是简单的线性关系。本文对 Zadeh 定义的标准 S 型函数进行改造[7],用于求取样本的隶属度。由标准 S 型函数改造而成的隶属度函数形式为

$$\mu(d_i;a,b,c) = \begin{cases} 1 & d_i \leqslant a \\ 1-2[(d_i-a)/(c-a)]^2 & a \leqslant d_i \leqslant b \\ 2[(d_i-c)/(c-a)]^2 & b \leqslant d_i \leqslant c \\ 0 & d_i \geqslant c \end{cases} \tag{3}$$

其中,d_i 为样本与所在类中心之间的距离,由式(1)确定。参数 a,b 是一个预先定义的参数,$b = \dfrac{(a+c)}{2}$,此时当 $d_i = b$ 时,$\mu(b;a,b,c) = 0.5$。

2.2　基于紧密度的隶属度函数

支持向量机方法中,最优分类面主要由支持向量决定,支持向量位于类边缘,而野值或含噪声的样本常常也位于类边缘附近,如果在确定样本隶属度时,将有效样本与野值或含噪声样本同等看待,则求出的分类面不是真正的最优分类面。前面介绍的两种确定隶属度的方法中,都是依据样本到类中心之间距离确定样本的隶属度,对类中每个样本都按照同一种方式进行考虑,对有效样本与含噪声的样本或野值无法区分开,因此,它们不能有效地反映样本的不确定性。

如图 1 所示为两个不同类中样本之间紧密度的差别。图 1(a)与(b)中,样本 $\boldsymbol{x}$ 到各自所在类中心之间的距离相等,如果仅依据距离来确定隶属度,则两者属于各自类的隶属度相同。然而,没有考虑图 1(a)中样本 $\boldsymbol{x}$ 与类中其他样本之间的距离远小

于图 1(b)中样本 x 与类中其他样本之间的距离这一实际情况，图 1(a)中样本 x 可能为有效样本，而图 1(b)中样本 x 为野值的可能性非常大。事实上，图 1(a)中样本 x 属于所在类的隶属度应大于图 1(b)中样本 x 属于所在类的隶属度。

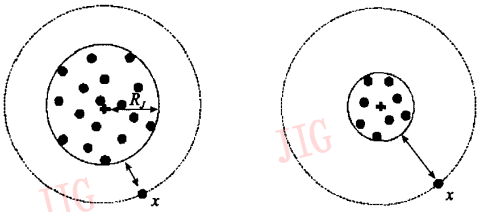

图 1 两个不同类中样本之间紧密度的差别示意图

Fig. 1 The difference of affinity among samples from two classes

针对这种情况，提出了基于样本紧密度的隶属度确定方法，在确定样本的隶属度时，既要考虑样本到所在类中心之间的距离，还要考虑样本与类中其他样本之间的关系，而样本与类中其他样本之间的关系可通过类中样本的紧密度来反映。

由以上分析可知，基于紧密度的隶属度的计算由两部分构成：

$$\mu(\boldsymbol{x}_i) = f(\mu_d(\boldsymbol{x}_i), \mu_k(\boldsymbol{x}_i, \bar{\boldsymbol{x}})) \tag{4}$$

其中，

(1)$\mu(\boldsymbol{x}_i)$为样本 $\boldsymbol{x}_i$ 属于所在类的隶属度。

(2)$\mu_d(\boldsymbol{x}_i)$由式(3)确定，反映样本 $\boldsymbol{x}_i$ 到所在类中心之间的距离关系。

(3)$\mu_k(\boldsymbol{x}_i, \bar{\boldsymbol{x}})$为样本 $\boldsymbol{x}_i$ 与所在类中心之间的模糊连接度[8]，反映 $\boldsymbol{x}_i$ 样本与类中其他样本之间的紧密度关系，其由下式确定：

$$\mu_k(\boldsymbol{x}_i, \bar{\boldsymbol{x}}) = \max_{\rho(\boldsymbol{x}_i, \bar{\boldsymbol{x}}) \in P(\boldsymbol{x}_i, \bar{\boldsymbol{x}})} [\min(\mu_k(\boldsymbol{c}_1, \boldsymbol{c}_2), \mu_k(\boldsymbol{c}_2, \boldsymbol{c}_3), \cdots, \mu_k(\boldsymbol{c}_{m-1}, \boldsymbol{c}_m))] \tag{5}$$

其中，$\rho(\boldsymbol{x}_i, \bar{\boldsymbol{x}})$表示从 $\boldsymbol{x}_i$ 到 $\bar{\boldsymbol{x}}$ 的一条路径，路径的各点用 $\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_m$ 表示，$m \geqslant 2$，$\boldsymbol{c}_1 = \boldsymbol{x}_i$，$\boldsymbol{c}_m = \bar{\boldsymbol{x}}_i$，$P(\boldsymbol{x}_i, \bar{\boldsymbol{x}})$表示从 $\boldsymbol{x}_i$ 到 $\bar{\boldsymbol{x}}$ 的所有路径的集合。

(4)$f(.,.)$表示为某种函数关系，本文取为乘积关系，这样式(4)就变为

$$\mu(\boldsymbol{x}_i) = \mu_d(\boldsymbol{x}_i) \times \mu_k(\boldsymbol{x}_i, \bar{\boldsymbol{x}}) \tag{6}$$

其中，k 为常数，主要是确保隶属度满足 $\mu(\boldsymbol{x}_i) \in (0,1]$。

3 模糊支持向量机

在采用模糊支持向量机进行分类时，相对于常规支持向量机的训练样本，除了样本的特征与类属标识外，模糊支持向量机训练的每个样本还增加了隶属度一项。

设训练样本集表示为$(y_1, \boldsymbol{x}_1, \mu(\boldsymbol{x}_1)), \cdots, (y_n, \boldsymbol{x}_n, \mu(\boldsymbol{x}_n))$

每个样本的特征表示为 $\boldsymbol{x}_i \in \boldsymbol{R}^N$，类标识为 $y_i \in \{-1, 1\}$，隶属度为 $0 < \mu(\boldsymbol{x}_i) \leqslant 1$。假设 $z = \varphi(\boldsymbol{x})$为将训练样本从原始模式空间 $\boldsymbol{R}^N$ 映射到高维特征空间 $\boldsymbol{Z}$ 之间的映射关系 φ。

由于隶属度 $\mu(\boldsymbol{x}_i)$表示该样本属于某类的可靠程度，ξ_i 是支持向量机目标函数中的分类误差项，则 $\mu(\boldsymbol{x}_i)\xi_i$ 为带权的误差项，由文献[1]得到最优分类面为下面目标函数的最优解。

$$\Phi(\boldsymbol{w}, \xi) = \frac{1}{2} \| \boldsymbol{w} \|^2 + C\left(\sum_{i=1}^{n} \mu(\boldsymbol{x}_i)\xi_i \right) \tag{7}$$

约束条件为

$$y_i[(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{z}_i) + b] - 1 + \xi_i \geqslant 0, i = 1, \cdots, n \tag{8}$$

$$\xi_i \geqslant 0, i = 1, \cdots, n \tag{9}$$

其中，惩罚因子 C 为常数，$\boldsymbol{w}$ 表示线性分类函数 y_i 的权系数。由式(7)可以看出，当 $\mu(\boldsymbol{x}_i)$ 很小时，减小了 ξ_i 在式(7)中的影响，以致于将相应的 $\boldsymbol{x}_i$ 看作不重要的样本。

而相应的最优分类面的判别函数式为

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{\boldsymbol{x}_i \in SV} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right) \tag{10}$$

其中，$K(\boldsymbol{x}_i, \boldsymbol{x})$称为核函数，$K(\boldsymbol{x}_i, \boldsymbol{x})$将高维特征空间中内积运算转化为低维模式空间上一个简单的函数计算。本文采用已在许多领域获得成功应用的高斯核 $K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\| \boldsymbol{x} - \boldsymbol{z} \|^2}{2\sigma^2} \right)$，其中，$\sigma$ 为高斯分布的宽度。

α_i 的条件式变为下式：

$$0 \leqslant \alpha_i \leqslant \mu(\boldsymbol{x}_i)C, \quad i = 1, \cdots, n \tag{11}$$

$\alpha_i > 0$ 相应的样本 $\boldsymbol{x}_i$ 为支持向量，这里有两种类型的支持向量，一种满足 $0 < \alpha_i < \mu(\boldsymbol{x}_i)C$ 的支持向量 $\boldsymbol{x}_i$ 位于分类面附近；另一种满足 $\alpha_i = \mu(\boldsymbol{x}_i)C$ 的支持向量 $\boldsymbol{x}_i$ 为错误分类样本。模糊支持向量机方法与支持向量机方法的差别在于，由于在模糊支持向量机中含有隶属度 $\mu(\boldsymbol{x}_i)$，同样 α_i 值的样本 $\boldsymbol{x}_i$ 在两种方法中可能属于不同类型的支持向量。

在支持向量机方法中参数 C 是一个自定义的惩罚因子，它控制对错分样本惩罚的程度，用来控制样本偏差与机器推广能力之间的平衡。C 越大，惩

罚就越大,对错分样本的约束程度就越大,得到分类面的间隔越小;随着 C 的降低,支持向量机忽略更多的样本,得到较大边缘间隔的分类面。在模糊支持向量机中,设置 C 为一个较大的值,如果取所有的隶属度 $\mu(x_i)=1$,则与支持向量机方法一样,容许更小的误分率,得到较窄边缘的分类面。在模糊支持向量机方法中,通过对不同样本赋予不同的隶属度 $\mu(x_i)$,达到对不同的样本采用不同程度的惩罚作用。更小隶属度 $\mu(x_i)$ 的样本 x_i 在训练中起更小的作用。

4　实验结果及分析

将支持向量机方法应用于医学图像分类与识别,由于医学图像具有模糊的特点,致使样本存在噪声及不确定性的问题,如果仍采用常规的支持向量机方法进行分类,则影响其分类精度。本文采用模糊支持向量机方法进行脑组织分类的实验。

采用在线图像库[9],体图像大小为 181 × 217 × 181,每片图像的厚度为 1mm,T1 加权的 MRI 图像。本实验采用含有 9% 的噪声及 40% 的灰度非均匀性的 22#与 32#切片。训练样本数为 1 500,图像特征采用文献[10]中采用的图像纹理与灰度特征。为了更好地评价基于紧密度的隶属度的模糊支持向量机方法的分类性能及抗噪能力,在采用模糊支持向量机方法时,同时分别采用基于线性距离的隶属度函数、基于 S 型函数的隶属度函数,同时还采用传统支持向量机方法进行了脑组织分类的对比实验,各种方法的分类错误率如表 1 所示。

表 1　模糊支持向量机与支持向量机方法分类错误率对比表

Tab. 1　The classification error rates of SVM and FSVM

切片号	传统支持向量机方法	模糊支持向量机方法			参数
		基于线性距离的隶属度	基于 S 型函数隶属度	基于紧密度的隶属度	
32#	13.8	13.5	11.6	9.6	$\sigma^2=0.5$
22#	9.82	9.9	9.78	8.5	$\sigma^2=10$

由表 1 可以看出,一般来说,采用模糊支持向量机方法,其分类错误率比采用支持向量机方法的错误率低,在使用的 3 种隶属度函数中,采用基于紧密度隶属度的模糊支持向量机方法抗噪性能最好,分类性能最强,例如 32#切片,其错误率由支持向量机方法的 13.8% 降为 9.6%。同时也可看到,不合适的隶属度会使模糊支持向量机的分类性能下降,如当对 22#切片进行分类,采用基于线性距离的隶属度时,其错误率比支持向量机方法的错误率还要高。图 2 为 32#切片中脑白质类样本的 3 种隶属度曲线分布图。从图 2 可以看出,基于紧密度的隶属度与样本到类中心的距离之间不是一一对应的关系,因为,在确定基于紧密度的隶属度时,既要考虑样本到所在类中心之间的距离,还要考虑样本与类中其他样本之间的关系,有利于将含噪声或野值样本从有效样本集中区分出来。

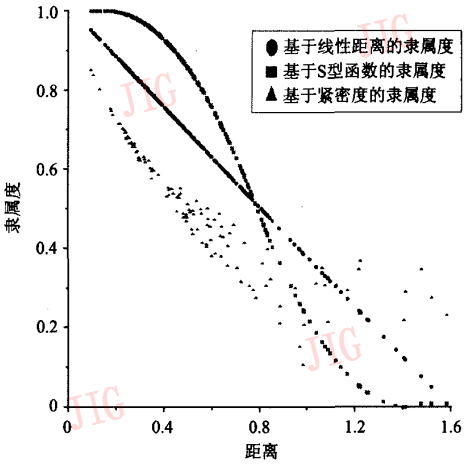

图 2　32#切片脑白质类样本的 3 种隶属度曲线分布图

Fig. 2　Three different fuzzy membership curves of brain white matter tissue of the 32nd slice

5　结　论

针对支持向量机方法中存在对噪声和野值敏感的问题,本文研究和分析了目前针对这些问题所提出的模糊支持向量机方法,而在模糊支持向量机方法中,隶属度函数的设计是整个模糊算法的关键所在,本文主要对隶属度函数进行了深入地研究,提出了一种基于类中样本紧密度的隶属度函数确定方法。即在确定样本的隶属度时,既要考虑样本到所在类中心之间的距离,还要考虑样本与类中其他样本之间的距离,而样本与类中其他样本之间的距离可通过类中样本的紧密度来反映。通过医学图像脑组织分类实验,采用模糊支持向量机方法,其分类错误率比采用支持向量机方法的错误率低,在使用的 3 种隶属度函数中,采用基于紧密度隶属度的模糊

支持向量机方法抗噪性能最好,分类性能最强。

参考文献(References)

1 Lin C F, Wan Sh D. Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks, 2002,**13**(2):464 ~ 471.

2 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.

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

4 Huang H P, Liu Y H. Fuzzy support vector machines for pattern recognition and data mining [J]. Internation Journal of Fuzzy Systems,2002,**4**(3):826 ~ 835.

5 Zhang J S, Leung Y W. Robust clustering by pruning outliers[J]. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics,2003,**33**(6):983 ~ 999.

6 George K, Dimitrios G, Nick K, *et al*. Efficient biased sampling for approximate clustering and outlier detection in large data sets[J]. IEEE Transactions on Knowledge and Data Engineering. 2003, **15**(5):1170 ~ 1187.

7 Bian Zhao-qi, Zhang Xue-gong. Pattern Recognition[M]. Beijing: Tsinghua University Press, 2000:135 ~ 136. [边肇祺,张学工编著. 模式识别[M]. 北京:清华大学出版社,2000:135 ~ 136.]

8 UduPa J K, Samarasekera S. Fuzzy connectedness and object definition: theory, algorithms, and applications in image segmentation[J]. Graphical Model and Image Processing, 1995, **58**(3): 246 ~ 261.

9 Cocosco C A, Kollokian V R, Kwan K S, *et al*. BrainWeb: online interface to a 3D MRI simulated brain database[J]. NeuroImage, 1997,**5**(4):425.

10 Zhang Xiang, Tian Jin-wen, Xiao Xiao-ling, *et al*. Support vector machine and its application in medical images classification[J]. Signal Processing. 2004,**20**(2):208 ~ 212. [张翔,田金文,肖晓玲等. 支持向量机及其在医学图像分类中的应用[J]. 信号处理, 2004,**20**(2):208 ~ 212.]