ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.5, May 2006, pp.951−958
DOI: 10.1360/jos170951

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于样本之间紧密度的模糊支持向量机方法*

张 翔[1,2+], 肖小玲[3], 徐光祐[1]

[1](清华大学 计算机科学与技术系,北京 100084)

[2](长江大学 地球物理与石油资源学院,湖北 荆州 434023)

[3](武汉理工大学 计算机科学与技术学院,湖北 武汉 430063)

Fuzzy Support Vector Machine Based on Affinity Among Samples

ZHANG Xiang[1,2+], XIAO Xiao-Ling[3], XU Guang-You[1]

[1](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2](School of Geophysics and Oil Resources, Yangtze University, Jingzhou 434023, China)

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

+ Corresponding author: Phn: +86-10-62782406, E-mail: xiang-zhang@tsinghua.edu.cn, http://www.tsinghua.edu.cn

Zhang X, Xiao XL, Xu GY. Fuzzy support vector machine based on affinity among samples. *Journal of Software*, 2006,17(5):951−958. http://www.jos.org.cn/1000-9825/17/951.htm

Abstract: Since SVM is very sensitive to outliers and noises in the training set, a fuzzy support vector machine algorithm based on affinity among samples is proposed in this paper. 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 affinity among samples. A method defining the affinity among samples is considered using a sphere with minimum volume while containing the maximum of the samples. Then, the fuzzy membership is defined according to the position of samples in sphere space. Compared with the fuzzy support vector machine algorithm based on the relation between a sample and its cluster center, this method effectively distinguishes between the valid samples and the 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 the fuzzy support vector machines based on the distance of a sample and its cluster center.

Key words: fuzzy support vector machine; affinity; classification

摘 要: 针对传统支持向量机方法中存在对噪声或野值敏感的问题,提出了一种基于紧密度的模糊支持向量机方法.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.通过样本之间的紧密度来描述类中各个样本之间的关系,利用包围同一类中样本的最小球半径大小来度量样本之间的紧密度.样本的隶属度依据样本在球中的位置,按照不同的规律确定.与基于样本与类中心之间关系构建的模糊支持向量机

* Supported by the National Natural Science Foundation of China under Grant No.60273005 (国家自然科学基金); the Chinese Postdoctoral Science Foundation under Grant No.2005038310 (中国博士后科学基金); the Natural Science Foundation of Hubei Province of China under Grant No.2004ABA043 (湖北省自然科学基金); the Key Science Technology Research Project of Hubei Provincial Department of Education under Grant No.D200612002 (湖北省教育厅科学技术研究重点项目)

Received 2005-09-24; Accepted 2005-11-08

方法相比,该方法有利于将野值或含噪声样本与有效样本进行区分.实验结果表明,与传统支持向量机方法及基于样本与类中心之间关系的模糊支持向量机方法相比,基于紧密度的模糊支持向量机方法具有更好的抗噪性能及分类能力.

关键词: 模糊支持向量机;紧密度;分类
中图法分类号: TP18 文献标识码: A

支持向量机(support vector machine,简称 SVM)被看作是对传统分类器的一个好的替代,特别是在高维数据空间下,具有较好的泛化能力[1],已在许多模式识别中得到了成功的应用[2].尽管支持向量机方法具有较好的推广能力,但由于在构造最优分类面时所有的样本具有相同的作用,因此,当训练样本中含有噪声或野值样本时,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.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,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,通常需要我们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[4].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.其主要原因在于: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.针对这种情况,本文提出了一种基于紧密度的隶属度确定方法,在确定样本的隶属度时,不仅要考虑样本与所在类中心之间的距离,还要考虑类中样本之间的紧密度,并以此构造基于紧密度的模糊支持向量机方法.

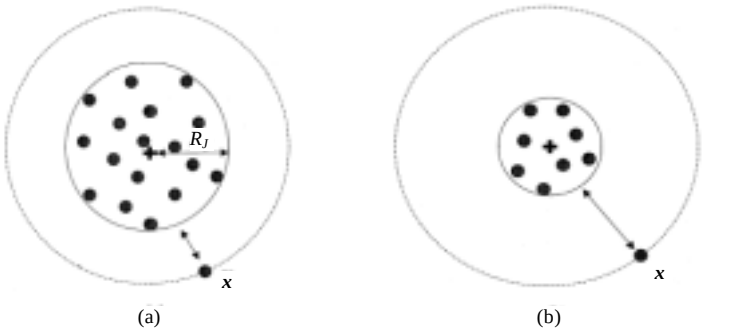
1 构造基于紧密度的隶属度函数

由于在支持向量机方法中,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,在确定样本隶属度时,如果无法将有效样本与野值或含噪声样本进行正确的区分,则求出的分类面不是真正的最优分类面.因此,在构建模糊支持向量机方法中,隶属度函数的设计非常关键,要求隶属度函数必须能够客观、准确地反映系统的不确定性.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.在文献[4]采用的基于距离的隶属度函数中,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线性函数.文献[6]使用 Zadeh 定义的 S 型函数隶属度函数,将样本的隶属度与样本到所在类中心的距离之间不再看作是简单的线性关系,而是一个非线性关系.在这两种确定隶属度的方法中,对类中每个样本都按照样本与类中心之间距离的准则进行考虑.然而,这种方式往往无法将有效样本与野值或含噪声的样本加以区分.

在图 1(a)与图 1(b)中,样本 $\boldsymbol{x}$ 到各自所在类中心之间的距离相等,如果仅依据距离的角度来确定隶属度,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 $\boldsymbol{x}$ 与类中其他样本之间的距离这一实际情况.图 1(a)中样本 $\boldsymbol{x}$ 可能为有效样本,而图 1(b)中样本 $\boldsymbol{x}$ 为野值的可能性非常大.事实上,图 1(a)中样本 $\boldsymbol{x}$ 属于所在类的隶属度应大于图 1(b)中样本 $\boldsymbol{x}$ 属于所在类的隶属度.因此,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,又要考虑样本与类中其他样本之间的距离.针对这种情况,本文提出将样本与类中其他样本之间的距离通过类中样本的紧密度来反映,并利用类中样本的紧密度来确定样本的隶属度函数的方法.

在确定样本的隶属度时,不仅要考虑样本与所在类中心之间的距离,还要考虑类中样本之间的紧密度.在此,借助文献[7]中样本聚类的思想,在特征空间中,可以用一个紧凑的球或超球将样本集包围起来.此时,样本之

间的紧密度可以通过包围样本集的最小球半径来度量.



(a) (b)

Fig.1 Difference of the affinity among samples at different two classes

图 1 两个不同类中样本之间紧密度的差别

1.1 最小球半径的确定

样本集的紧密度通过包围样本集的最小球半径来度量.在样本集数目固定的情况下,球半径越大,样本的紧密度越小;反之,球半径越小,样本的紧密度越大.因此,在基于样本紧密度的隶属度确定方法中,需要确定能够包围样本集的最小球半径.当样本集中不存在噪声或野值样本时,则寻找一个能够包围所有样本的最小球半径.当样本集中含有噪声或野值样本时,可以允许一小部分样本位于球的外面.此时,则寻找一个能够包围样本集中大多数样本的最小球半径.

设样本集中 n 个样本表示为$\{\boldsymbol{x}_i,i=1,2,\ldots,n\}$,当样本集中不存在噪声或野值样本,或者事先不知道样本集中是否存在噪声或野值样本时,通过引入一个非负松弛变量$\xi_i,i=1,2,\ldots,n$ 来允许一部分样本位于球的外面.采用寻找最优分类面类似的方法,通过对下面目标函数的最小化得到最小包围球.即表示为

$$\Phi(R,a,\xi)=R^2+D\left(\sum_{i=1}^{n}\xi_i\right) \tag{1}$$

其中,R 为能够包围样本集的最小球半径;a 为球中心;$D>0$ 是一个自定义的惩罚因子,用来控制包围球的体积与允许球外面存在样本的个数之间的折衷.D 越大,惩罚就越大,对允许球外面存在样本的约束程度也就越大.

约束条件为

$$||\boldsymbol{x}_i-a||^2\le R^2+\xi_i,\ i=1,\ldots,n \tag{2}$$

$$\xi_i\ge 0,\ i=1,\ldots,n \tag{3}$$

为了求解带约束条件式(2)与式(3)的优化问题式(1),可以定义如下的 Lagrange 函数:

$$L(R,a,\beta,\gamma,\xi)=R^2+D\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\beta_i\left\{R^2+\xi_i-\left\|\boldsymbol{x}_i-a\right\|^2\right\}-\sum_{i=1}^{n}\gamma_i\xi_i \tag{4}$$

其中,$\beta_i\ge 0$ 为 Lagrange 系数.

求解式(4)的最小值,可以令该泛函对 R,a 及ξ_i 求偏导,并令它们等于 0,得到

$$\sum_i\beta_i=1 \tag{5}$$

$$a=\sum_i\beta_i\boldsymbol{x}_i \tag{6}$$

$$D-\beta_i-\gamma_i=0,\ i=1,\ldots,n \tag{7}$$

将式(4)展开并进行组合为

$$L(R,a,\beta,\gamma,\xi)=R^2\left(1-\sum_{i=1}^{n}\beta_i\right)+\sum_{i=1}^{n}\xi_i(D-\beta_i-\gamma_i)+\sum_{i=1}^{n}\beta_i\left\{\left\|\boldsymbol{x}_i-a\right\|^2\right\} \tag{8}$$

将约束条件式(5)、式(6)及式(7)代入式(8)中,并进行合并整理,得到

$$L(R,a,\beta,\gamma,\xi)=\sum_{i=1}^{n}\beta_i(\boldsymbol{x}_i\cdot\boldsymbol{x}_i)-2a\sum_{i=1}^{n}\beta_i\boldsymbol{x}_i+a^2\sum_{i=1}^{n}\beta_i \tag{9}$$

参考式(5),将式(6)变为

$$a\sum_{i=1}^{n}\beta_i=\sum_{i=1}^{n}\beta_i\boldsymbol{x}_i \tag{10}$$

将式(10)代入式(9),即得

$$Q(\beta)=L(R,a,\beta,\gamma,\xi)=\sum_{i=1}^{n}\beta_i(\boldsymbol{x}_i\cdot\boldsymbol{x}_i)-\sum_{i,j=1}^{n}\beta_i\beta_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j) \tag{11}$$

即在约束条件

$$\sum_{i=1}^{n}\beta_i=1 \tag{12}$$

$$0\le\beta_i\le D,\ i=1,\ldots,n \tag{13}$$

之下对β_i求解式(11)函数的最大值.

通过求解式(10)二次规划问题,就可以得到样本对应的 Lagrange 系数$\beta_i,i=1,\ldots,n$.

由式(6)可知,最小包围球中心为带权系数β_i 的样本线性加权组合.当$\beta_i>0$ 时,对应的样本称为支持样本;$\beta_i=D$ 时,对应的样本位于包围球外边,称为野值或含噪声的样本.当 $0<\beta_i<D$ 时,对应的样本用来描绘包围球,位于包围球附近.因此,最小包围球的半径由 $0<\beta_i<D$ 中对应的任意样本与球中心之间的距离来确定,即为

$$R=||\boldsymbol{x}_i-a|| \tag{14}$$

其中,$\boldsymbol{x}_i$为 Lagrange 系数 $0<\beta_i<D$ 中对应的任意支持样本.

1.2 基于紧密度隶属度的计算

样本之间的紧密度可以通过包围样本集的最小球半径来度量.因此,在确定基于紧密度的隶属度时,依据最小包围球半径来确定样本集中样本的隶属度.对分布在半径内、外的样本,分别采用两种不同的方式计算其各自样本的隶属度.基于紧密度的隶属度计算公式为

$$\mu(\boldsymbol{x}_i)=\begin{cases}0.6*\left(\dfrac{1-d(x_i)/R}{1+d(x_i)/R}\right)+0.4, & d(x_i)\le R\\ 0.4*\left(\dfrac{1}{1+(d(x_i)-R)}\right), & d(x_i)>R\end{cases} \tag{15}$$

其中,R 为样本集中最小包围球半径;$d(\boldsymbol{x}_i)$为样本集中样本 $\boldsymbol{x}_i$ 到其最小包围球中心 a 之间的距离.其计算公式为

$$d(\boldsymbol{x}_i)=||\boldsymbol{x}_i-a||,\ i=1,\ldots,n \tag{16}$$

由式(15)定义的基于紧密度的隶属度$\mu(\boldsymbol{x}_i)$可以看出:样本到最小包围球中心之间的距离越大,则该样本属于该样本集的隶属度就越小;同时,考虑了样本集中样本在特征空间中分布范围的影响,位于球半径内的样本,其隶属度都大于 0.4;而位于球半径外的样本,其隶属度最大值为 0.4.虽然样本的隶属度与其到最小包围球中心之间的距离成反比,但位于球半径内、外样本的隶属度与其到最小包围球中心之间距离的变化规律不同.由于野值或含噪声的样本位于最小包围球外边,因此,采用基于紧密度的隶属度确定方法,能够有效地将野值或含噪声的样本与样本集中有效样本区分开,并按不同的规律计算它们各自相应的隶属度.

2 构造基于紧密度的模糊支持向量机

在依据样本集的紧密度确定样本的隶属度之后,就可以构造基于紧密度的模糊支持向量机.在采用基于紧密度的模糊支持向量机进行分类时,相对于传统支持向量机的训练样本,除了样本的特征与类属标识以外,基于紧密度的模糊支持向量机训练的每个样本还增加了基于紧密度的隶属度一项.

设训练样本集表示为$(y_1,\boldsymbol{x}_1,\mu_1),\ldots,(y_n,\boldsymbol{x}_n,\mu_n)$,其中,$\mu_i,i=1,\ldots,n$ 为样本 $\boldsymbol{x}_i$ 属于所在类的基于紧密度的隶属度

$\mu(\boldsymbol{x}_i)$.

每个样本的特征表示为 $\boldsymbol{x}_i \in R^N$,类标识为 $y_i \in \{-1,1\}$,隶属度为 $0<\mu_i \le 1$.假设 $\boldsymbol{z}=\varphi(\boldsymbol{x})$为将训练样本从原始模式空间 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μ_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$\mu_i\xi_i$ 为带权的误差项,基于紧密度的模糊支持向量机的最优分类面为下面的目标函数的最优解:

$$\Phi(\boldsymbol{w},\xi)=\frac{1}{2}\|\boldsymbol{w}\|^2+C\left(\sum_{i=1}^{n}\mu_i\xi_i\right) \tag{17}$$

约束条件为

$$y_i[(\boldsymbol{w}^T\cdot\boldsymbol{z}_i)+b]-1+\xi_i\ge 0,\ i=1,\ldots,n \tag{18}$$

$$\xi_i\ge 0,\ i=1,\ldots,n \tag{19}$$

其中,惩罚因子 C 为常数.由式(17)可以看出:当μ_i 很小时,减小了ξ_i 在式(17)中的影响,以致将相应的 $\boldsymbol{x}_i$ 看作不重要的样本.

为了求解带约束条件的优化问题式(17),可以定义如下的 Lagrange 函数

$$L(\boldsymbol{w},b,\xi,\alpha,\beta)=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\mu_i\xi_i-\sum_{i=1}^{n}\alpha_i\{y_i[(\boldsymbol{w}^T\boldsymbol{z}_i)+b]-1+\xi_i\}-\sum_{i=1}^{n}\beta_i\xi_i \tag{20}$$

其中,$\alpha_i\ge 0$ 为 Lagrange 系数.

求解式(20)的最小值,可以令该泛函对 $\boldsymbol{w}$,b 及ξ_i 求偏导,并令它们等于 0

$$\frac{\partial L(\boldsymbol{w},b,\xi,\alpha,\beta)}{\partial \boldsymbol{w}}=\boldsymbol{w}-\sum_{i=1}^{n}\alpha_i y_i\boldsymbol{z}_i=0 \tag{21}$$

$$\frac{\partial L(\boldsymbol{w},b,\xi,\alpha,\beta)}{\partial b}=-\sum_{i=1}^{n}\alpha_i y_i=0 \tag{22}$$

$$\frac{\partial L(\boldsymbol{w},b,\xi,\alpha,\beta)}{\partial \xi_i}=\mu_i C-\alpha_i-\beta_i=0 \tag{23}$$

将式(21)、式(22)及式(23)代入式(20),则基于紧密度的模糊支持向量机的最优分类面问题转化为较为简单的对偶问题,即在约束条件

$$\sum_{i=1}^{n}y_i\alpha_i=0 \tag{24}$$

$$0\le\alpha_i\le\mu_i C,\ i=1,\ldots,n \tag{25}$$

之下对α_i 求解下列函数的最大值:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i\cdot\boldsymbol{x}_j) \tag{26}$$

通过对式(26)二次规划问题的求解,求得各样本对应的 Lagrange 系数α_i,$i=1,\ldots,n$,则基于紧密度的模糊支持向量机方法的最优判别函数为

$$f(\boldsymbol{x})=\mathrm{sgn}\{(\boldsymbol{w}\cdot\boldsymbol{x})+b\}=\mathrm{sgn}\left\{\sum_{\boldsymbol{x}_i\in SV}\alpha_i y_i K(\boldsymbol{x}_i\cdot\boldsymbol{x})+b\right\} \tag{27}$$

其中,Lagrange 系数α_i 满足条件:

$$0<\alpha_i\le\mu_i C \tag{28}$$

基于紧密度的模糊支持向量机方法与传统支持向量机方法相比,最终得到的最优判别函数与传统支持向量机方法得到的最优判别函数几乎完全相同,只是在条件式(28)中增加了隶属度μ_i.

$\alpha_i>0$ 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 $0<\alpha_i<\mu_i C$ 的支持向量 $\boldsymbol{x}_i$ 位于分类面附近;另一种满足$\alpha_i=\mu_i C$ 的支持向量 $\boldsymbol{x}_i$ 为错误分类样本.基于紧密度的模糊支持向量机方法与传统支持向量机方法的差别在于:由于在基于紧密度的模糊支持向量机中含有隶属度μ_i,相同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,惩罚越大,对错分样本的约束程度就越大,得到分类面的间隔就越小;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.在基于紧密度的模糊支持向量机中,我们设置 C 为一个较大的值,如果取所有的隶属度 μ_i=1,则与传统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.在基于紧密度的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 μ_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 μ_i 的样本 $\boldsymbol{x}_i$ 在训练中起更小的作用.

3 实验结果

3.1 含野值样本的实验结果

为了展示样本集中含有野值样本对支持向量机方法的影响以及基于紧密度的模糊支持向量机方法优良的分类能力,本节采用传统支持向量机方法与基于紧密度的模糊支持向量机方法,对含有野值的样本的两类样本分别进行分类对比实验.图 2 为传统支持向量机方法分类结果;基于紧密度的模糊支持向量机的分类结果如图 3 所示,图中正方形符号代表一类样本,十字形符号代表另一类样本.

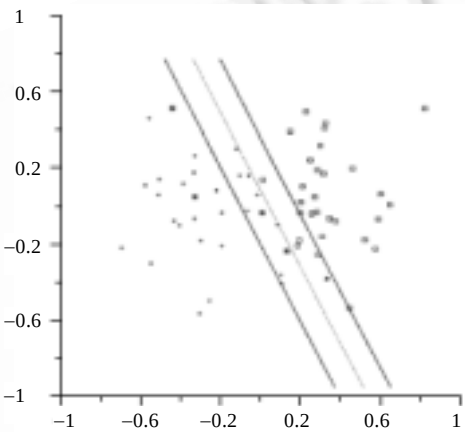


Fig.2 The result for the traditional SVM

图 2 传统支持向量机方法结果

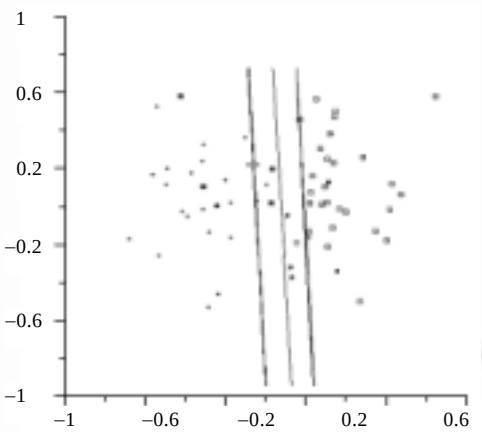


Fig.3 The result for the FSVM based on the affinity

图 3 基于紧密度的模糊支持向量机方法结果

由图 2 和图 3 可以看出:采用传统支持向量机方法与基于紧密度的模糊支持向量机方法进行样本的分类,由于在两类中都存在少数野值样本,致使两种方法的最优分类面完全不同.由于野值样本的存在,严重地影响了传统支持向量机方法的最优分类面;在基于紧密度的模糊支持向量机方法中,有效地识别了野值样本,并对野值样本与有效样本按照不同规律赋予相应的隶属度,使野值样本在构造最优分类面时起较小的作用,从而使基于紧密度的模糊支持向量机不受样本集中野值样本的影响,具有更好的抗野值样本的能力.

3.2 医学图像中脑组织分类实验结果

医学图像具有复杂性和多样性的特点.由于组织本身的特性差异,而且医学图像的形成受到诸如噪音、场偏移效应、局部体效应和组织运动等的影响,医学图像与普通图像比较,不可避免地具有模糊、不均匀性等特点.为了更好地评价基于紧密度的模糊支持向量机方法的分类性能及抗噪能力,我们采用来自 McGill 大学 McConnell 脑图像中心的在线图像库[8],将基于紧密度的模糊支持向量机方法进行脑组织分类的实验.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#,32#与 34#切片.训练样本数为 1 500,图像特征采用图像纹理与灰度特征[9].在采用基于紧密度的模糊支持向量机方法进行实验时,同时分别采用基于线性距离的模糊支持向量机方法、基于 S 型函数的模糊支持

向量机方法及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见表 1.

Table 1 The comparative results for classification error rate among the different SVM methods

表 **1** 几种不同支持向量机方法分类错误率对比

The slice No.	The classification error rate (%)			
	The traditional SVM	The FSVM based on the linear function	The FSVM based on the *S* shape function	The FSVM based on the affinity among samples
22#	9.82	9.9	9.78	8.75
32#	13.8	13.5	11.6	9.8
34#	12.07	9.46	8.68	7.83

由表 1 可以看出:当图像中含有噪声时,采用模糊支持向量机方法,其分类错误率比采用传统支持向量机方法的错误率要低.在 3 种模糊支持向量机方法中,采用基于紧密度的模糊支持向量机方法抗噪性能最好,分类性能最强.例如对 34#切片来看,其错误率由支持向量机方法的 12.07%降为 7.83%.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模糊支持向量机方法时,其错误率比传统支持向量机方法的错误率还要高.

图 4 为 34#切片原始图像,图 5 与图 6 分别为采用传统支持向量机方法及基于紧密度的模糊支持向量机方法对 34#切片进行脑组织分类的结果.从图 5 与图 6 的对比可以看出:由于原始图像中含有大量的噪声,传统支持向量机方法的分类结果严重受噪声的影响;而采用本文提出的基于紧密度的模糊支持向量机方法,较好地将脑组织进行了分类,具有较好的抗噪性能.

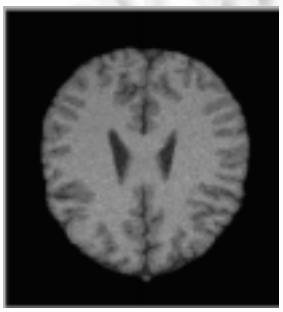


Fig.4 The original image for the 34th slice

图 4 34#切片原始图像

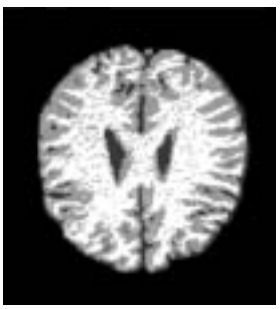


Fig.5 The results for the traditional SVM for the 34th slice

图 5 34#切片传统支持向量机分割结果

Fig.6 The results for the FSVM based on the affinity for the 34th slice

图 6 34#切片基于紧密度的模糊支持向量机分割结果

4 结束语

针对传统支持向量机方法中存在对噪声与野值敏感的问题,本文研究了一种基于紧密度的模糊支持向量机方法.在确定样本的隶属度时,不仅考虑了样本与所在类中心之间的距离,还考虑了样本集中样本之间的紧密度.在特征空间中,可以用一个紧凑的球或超球将样本集包围起来.此时,样本之间的紧密度可以通过包围样本集的最小球半径来度量.对分布在半径内、外的样本,分别采用两种不同的方式计算其各自样本的隶属度.由于考虑了样本之间的紧密度,因此,基于紧密度的隶属度相对于基于距离的隶属度的确定方法能够更有效地将野值或含噪声的样本与样本集中有效样本进行区分,并按不同的规律赋值它们各自相应的隶属度,从而更好地反映样本在基于紧密度的模糊支持向量机目标函数中所起的作用.

为了更好地评价基于紧密度的模糊支持向量机方法的分类性能及抗噪能力.使用仿真 MR 图像数据,并与基于线性距离的模糊支持向量机方法、基于 S 型函数的模糊支持向量机方法及传统支持向量机方法进行了脑组织分类的对比实验.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传统支持向量机方法的错误率要低;而本文提出的基于紧密度的模糊支持向量机方法相对于另外两种模糊支持向量机方法,具有更好的

抗噪性能及分类能力.

References:

[1] Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag, 1995.

[2] Burges C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998,2(2): 121−167.

[3] Lin CF, Wan SD. Fuzzy support vector machines. IEEE Trans. on Neural Networks, 2002,13(2):464−471.

[4] Huang HP, Liu YH. Fuzzy support vector machines for pattern recognition and data mining. Int'l Journal of Fuzzy Systems, 2002, 4(3):826−835.

[5] Chiang JH, Hao PY. A new kernel-based fuzzy clustering approach: Support vector clustering with cell growing. IEEE Trans. on Fuzzy Systems, 2003,11(4):518−527.

[6] Bian ZQ, Zhang XG, *et al*. Pattern Recognition. Beijing: Tsinghua University Press, 2000 (in Chinese).

[7] David MJ, Robert PW. Support vector domain description. Pattern Recognition Letters, 1999,20:1191−1199.

[8] Cocosco CA, Kollokian V, Kwan RKS, Evans AC. BrainWeb: Online interface to a 3D MRI simulated brain database. NeuroImage, 1997,5(4):425.

[9] Zhang X, Tian JW, Xiao XL, Liu J. Support vector machine and its application in medical images classification. Signal Processing, 2004,20(2):208−212 (in Chinese with English abstract).

附中文参考文献:

[6] 边肇祺,张学工,等.模式识别.北京:清华大学出版社,2000.

[9] 张翔,田金文,肖晓玲,柳健.支持向量机及其在医学图像分类中的应用.信号处理,2004,20(2):208−212.

张翔(1969－),男,湖北蕲春人,博士,副教授,主要研究领域为图像处理,计算机视觉,模式识别.

肖小玲(1973－),女,博士生,副教授,主要研究领域为计算机网络,模式识别.

徐光祐(1940－),男,教授,博士生导师,CCF高级会员,主要研究领域为计算机视觉,普适计算.