文章编号:1000-582X(2012)08-127-07

后验概率加权的模糊隶属度函数

魏　延[1,2],李晓虹[1],邬　啸[1]
(1.重庆师范大学 计算机与信息科学学院,重庆 401331;2.重庆大学 自动化学院,重庆 400044)

摘　要:模糊支持向量机(FSVM)中的模糊隶属度函数确定一直是一个难点问题。针对支持向量分类机对噪声数据或孤立点非常敏感的问题,受贝叶斯决策理论的启发,结合样本密度特性,研究样本点相对于同类和异类的关系,对各样本点分布的紧密程度给出了描述,构造了样本点的后验概率与样本密度的加权方法,提出了一种新的加权模糊隶属度函数构造。该方法避免了对噪声数据和孤立点的检测。通过建立基于提出模糊隶属函数的 FSVM 进行仿真,实验表明,提出的模糊隶属度函数构造的后验概率加权方法的有效性。

关键词:隶属度函数;模糊集;后验概率;样本密度;模糊支持向量机
中图分类号:TP274　　**文献标志码**:A

Design fuzzy membership functions based on the posterior probability weighting

WEI Yan[1,2], *LI Xiaohong*[1], *WU Xiao*[1]
(1. College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, P. R. China; 2. College of Automation, Chongqing University, Chongqing 400044, P. R. China)

Abstract: The determination of fuzzy membership function in the fuzzy support vector machine (FSVM) is a difficult problem. To solve the problem of being sensitive to the noises and outliers in support vector machine, by the inspiration of Bayesian decision theory, combining with sample density characteristics, sample points relation between same class and other class is researched, and the tightness on each sample points is described. Based on that, method of posterior probability and sample density weight are given to each sample, and new fuzzy membership function is proposed. The detection of the noises and outliers is avoided by this method. Numerical simulation shows that the improved fuzzy membership function method is effective.

Key words: membership functions; fuzzy sets; posterior probability; sample density; fuzzy support vector machine

支持向量机(SVM)是一个强有力的机器学习方法,是建立在统计学习理论 VC 维和结构风险最小化原则基础上,其目的是寻找最优分类超平面,并具有良好的推广性能。然而在实际应用中,由于存在噪声数据及孤立点,使得数据样本对超平面的影响是不同的。尤其是对于具有不确定性的分类问

收稿日期:2012-02-20
基金项目:国家自然科学基金资助项目(60974090);重庆市教育委员会科学技术研究项目(KJ090823,KJ110629)
作者简介:魏延(1970-),男,重庆师范大学副教授,博士,主要从事机器学习与智能计算研究,(Tel)023-65910278;(E-mail)weiyan@cqnu.edu.cn。

题,样本不能明确地属于某一类,而是以某一程度隶属于某一类[1]。为解决这类带有不确定性的问题出现了模糊支持向量机(FSVM)。

在FSVM理论中,样本隶属度函数的设计是一个关键和难点问题。许多学者提出了相应的隶属度函数构造方法,Liu等人[2]提出了基于类中心距离和类属性的样本隶属度构造方法,并建立了具有代表性的FSVM方法;Huang等人[3]采用样本距离定义隶属度函数,将样本的隶属度看作是问题空间中样本点与其所在类中心之间距离的线性函数;文献[4]使用S型函数,将样本的隶属度与样本到所在类中心的距离之间看作是一种非线性关系;唐浩等人[5]结合k邻近法提出了样本点到类中心距离和对各样本点排列的紧密程度进行估计的隶属函数构造方法。这些隶属度函数的构造都是基于样本集中的距离,没有充分考虑样本数据实际存在的不确定性[6-7]。同时,依据样本点到类中心距离确定样本隶属度时,由于支持向量均位于两类样本的相对边界,距两类类中心的距离较远,若按照已有文献提出的减小孤立点和噪声数据作用的方法确定隶属度,在减小其作用的同时,也大大减小了支持向量对分类超平面的作用。受贝叶斯决策理论的启发,文献[8]利用后验概率来表述样本的不确定性,建立了后验概率支持向量机框架;文献[9-10]初步研究了基于后验概率加权的FSVM方法,但在样本隶属度确定中对孤立点单独处理,需要先检测孤立点。

事实上,FSVM中的样本隶属度不仅要求能准确、客观地反映系统存在的不确定性,而且要准确描述样本点在问题空间中的位置分布关系。笔者针对两分类模糊支持向量机,考虑样本的相对于异类的关系,结合贝叶斯理论和样本分布密度,提出一种基于后验概率加权的模糊隶属度确定方法,以避免对孤立点和噪声数据的检测。

1 贝叶斯决策规则

贝叶斯决策[4,11]就是在对分类样本集数据进行概率分析的基础上,生成分类决策规则,应用生成的决策规则进而对新数据依据概率的方法进行分类。贝叶斯决策有许多方法生成决策规则[4],在模式分类问题中,希望尽量减少样本数据分类的错误,基于此,利用贝叶斯公式,就能生成使错误率为最小的分类规则,称之为基于最小错误率的贝叶斯决策。

设分类数据样本 $x=(x_1,x_2,\cdots x_l)\in\mathbf{R}^n$,分类问题有 m 个类别,各类别状态用 $\omega_j(j=1,2,\cdots,m)$ 表示,$\omega_1,\omega_2,\cdots,\omega_m$ 可以看作是样本空间 $\mathbf{R}^n$ 的一个划分。在贝叶斯决策理论中,类别状态 ω_j 看作是一个随机量,其出现的概率 $P(\omega_j)$ 可以通过样本先验知识进行估计,称为先验概率。由于先验概率提供的分类信息量少,还须利用对数据样本 x 的信息,即类条件概率密度函数 $p(x|\omega_j)$ 进行分类决策。

由贝叶斯决策假设,对应于各个类别 ω_j 出现的先验概率 $P(\omega_j)$ 及类条件概率密度函数 $p(x|\omega_j)$ 是已知的,利用贝叶斯公式

$$P(\omega_j\mid x)=\frac{p(x\mid\omega_j)P(\omega_j)}{\sum_{i=1}^{m}p(x\mid\omega_i)P(\omega_i)},\qquad(1)$$

得到的条件概率 $P(\omega_j|x)$ 称为样本类别状态的后验概率,则基于最小错误率的贝叶斯决策规则描述为[4]

如果 $P(\omega_j\mid x)=\max_i P(\omega_i\mid x)$

则 $x\in\omega_j$,其中 $j\in\{1,2,\cdots,m\}$。

由此可见,贝叶斯决策是在观测到 n 维空间样本 x 发生的条件下,类别 ω_j 的所有条件概率中最大者为样本 x 应所属的类,这样可使分类决策错误率最小。

然而,贝叶斯决策规则的使用除要求分类类别数已知外,还有一个重要的前提,即各类别的先验概率以及类条件概率密度均为已知。实际上,先验概率和类条件概率密度一般情况是很难确定的。

2 样本集中的样本密度

2.1 样本密度的概念

对给定的数据样本集,通常并不能确切地给出样本集的数据分布模型,但就样本分布的稠密程度而言,如果某个样本点 x 周围的样本数越多,则该样本点 x 被认为分布越紧密,在该样本点处的样本分布密度就越大,反之相反。

对数据样本集

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\},$$

其中 $x_i\in\mathbf{R}^n$,$y_i\in\{-1,+1\}$,$i=1,2,\cdots,l$,计算 l 个样本点两两之间的距离 $d_{ij}=\|x_i-x_j\|$,形成样本集 T 的距离矩阵

$$\boldsymbol{R}=\begin{bmatrix}d_{11}&d_{12}&\cdots&d_{1l}\\d_{21}&d_{22}&\cdots&d_{2l}\\\vdots&\vdots&\vdots&\vdots\\d_{l1}&d_{l2}&\cdots&d_{ll}\end{bmatrix}。\qquad(2)$$

则所有样本点的平均距离 D 和样本间最大距离 d 分别为

$$D=\frac{1}{C_l^2}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l}d_{ij}=\frac{2}{l(l-1)}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l}d_{ij},\quad(3)$$

$$d = \max_{i,j=1,2,\cdots,l} d_{ij}。\tag{4}$$

利用样本集中样本点的平均距离和样本间最大距离,进而定义样本集的样本平均密度和样本密度。

定义 1[12]　设数据样本集 T 的维度为 n,样本的个数为 l,数据集中所有样本间最大距离为 d,则样本集 T 的样本平均密度为:

$$\rho = \frac{l}{t\left(\frac{\sqrt{3}}{2}\times d\right)^n}。\tag{5}$$

其中 t 为常系数,为简便取 $t=1$,此时(5)式右端分母表示包含样本集 T 中所有样本的 n 维超球空间的超"体积"。

定义 1 中,如果只利用数据集中所有样本间最大距离 d 作为直径形成超球,可能会因遗漏一些样本点,使得计算的平均密度不能反映客观实际情况。而适当扩大最大距离 d 之后的封闭区域则能包含所有的样本点,从几何学上扩大 d 的$\sqrt{3}$倍,恰好能够保证所形成封闭区域包含数据集中的所有样本点[12]。

为了进一步区别样本集 T 中样本点的位置关系,定义样本点的样本密度概念。考虑样本点的一封闭小邻域的样本分布,为与平均密度定义一致,在计算每个样本点 $x_i(i=1,2,\cdots,l)$的密度时,仍将样本邻域有效直径延长至原来的$\sqrt{3}$倍。同时,给定 λ $(0<\lambda<1)$,即以$\sqrt{3}\lambda D$ 为直径的封闭区域,扫描距离矩阵 R,计算每个样本点单位封闭区域的样本数 l_i,即与样本点 x_i 间距离小于$\sqrt{3}\lambda D/2$ 的样本点计数。

定义 2　样本集 T 中样本点平均距离为 D,给定 $\lambda(0<\lambda<1)$,距样本点 x_i 小于$\sqrt{3}\lambda D/2$ 的样本数为 l_i,则样本点 x_i 的样本密度定义为

$$\rho_i = \frac{l_i}{\left(\frac{\sqrt{3}}{2}\times\lambda D\right)^n}, i = 1,2,\cdots,l。\tag{6}$$

样本集的平均密度和样本密度类似于物理学上关于密度的定义,其示意如图 1 所示。

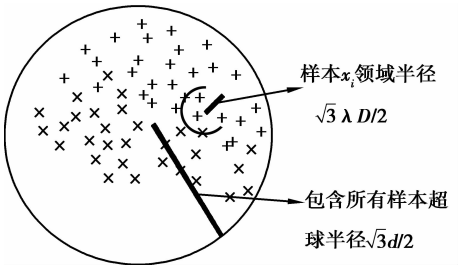

图 1　样本集平均密度与样本密度示意图

式(6)定义了样本集 T 中各个样本的密度,样本密度的不同,体现了样本点位置空间的稠密关系,样本密度 ρ_i 越大,表示样本 x_i 的邻近样本数越多,该样本点对支持向量机分类的作用越大;样本密度 ρ_i 越小,表示样本 x_i 的邻近样本数越少,该样本点对支持向量机分类的作用越小。

2.2　特征空间的样本密度

支持向量机在训练学习过程中,当把原始空间中的样本点通过 $\Phi(x)$映射到特征空间时,样本在特征空间中将重新分布。由于映射 $\Phi(x)$的具体形式未知,样本在特征空间中的分布也是未知的,所以不能确定原始空间中的噪声异常数据在特征空间里是否也同为噪声异常数据。

样本密度的定义主要是利用了原始空间样本点之间的距离,特征空间中样本点之间的距离可以通过核函数 $K(x,x')$来求得

$$d(\Phi(x),\Phi(x')) = \|\Phi(x)-\Phi(x')\| = \sqrt{[\Phi(x)-\Phi(x'),\Phi(x)-\Phi(x')]} = \sqrt{[\Phi(x),\Phi(x)]-2[\Phi(x),\Phi(x')]+[\Phi(x'),\Phi(x')]} = \sqrt{K(x,x)-2K(x,x')+K(x',x')}。\tag{7}$$

利用式(7),将距离矩阵 R 中的 d_{ij} 转换成特征空间中样本点的距离,得到特征空间中的距离矩阵,再利用式(5)和(6)加以计算,就可以得到特征空间中样本点的平均密度和样本密度。

3　基于后验概率加权的隶属度函数

在 FSVM 建模中,隶属度函数的构造是关键之一,目前已经提出了许多隶属度函数的构造方法。本文通过数据样本集的平均密度和样本密度,利用样本集的后验概率进行加权,设计出一种新的加权隶属度函数构造方法。

3.1　确定后验概率的经验性方法

由贝叶斯决策规则,若已知各样本点的类先验概率及类条件概率,由贝叶斯公式式(1)便可确定样本点的后验概率。然而在实际应用中,通常样本集类先验概率和样本点类条件概率是未知的。对两分类 FSVM,使用数据样本集后验概率的一种经验性求法[8],在给定样本集的基础上估计类先验概率以及样本点的类条件概率。同时考虑样本集中样本 x_i 邻域中同类样本数和异类样本数,进而准确描述样本点后验概率。

定义 3　用 l_j 表示属于类 $\omega_j(j=1,2)$的样本数,$l=l_1+l_2$ 为 FSVM 的训练样本集 T 中样本数,则基于数据样本集对类先验概率的估计为

$$\hat{P}(\omega_j)=\frac{l_j}{l},j=1,2。\tag{8}$$

事实上,若假设类 ω_j 的类先验概率为 $P_j(j=1,2)$,在样本独立同分布下,l 个样本中有 l_j 个样本属于类 ω_j 的概率是

$$P(l_j)=\left(\frac{l_j}{l}\right)P_j^{l_j}(1-P_j)^{l-l_j},j=1,2。$$

l_j 的数学期望为 $E(l_j)=P_jl\theta$,且二项分布在均值处具有陡峭的尖峰,因此比率 $\frac{l_j}{l}$ 是对类先验概率 P_j 的一个很好估计。

关于样本点 x_i 处的类条件概率密度,为避免估计整个样本空间的分布,用输入样本 x 落入 x_i 的一小邻域的条件概率来代替该处的类条件概率密度。

由式(2)~(3)知,数据样本集中样本点的平均距离为 D,由于样本数目 l 是有限的,D 存在且有界。选定 $0<\lambda<1$,以超球

$$S(x_i,\sqrt{3}\lambda D/2)=\{x\mid\|x-x_i\|<\sqrt{3}\lambda D/2\},$$

作为 x_i 邻域,如图1所示,设数据样本集中落入 $S(x_i,\sqrt{3}\lambda D/2)$ 的属于类 $\omega_j(j=1,2)$ 的样本数目为 k_i^j。

定义4 样本集 T 中样本点 x_i 的类条件概率密度的估计为

$$\hat{p}(x_i\mid\omega_j)=\frac{k_i^j}{l_j},j=1,2。\tag{9}$$

这样,有了类先验概率及类条件概率密度的经验估计,按照贝叶斯公式(1),就可以得到样本点 x_i 的后验概率

$$P(\omega_j\mid x_i)=\frac{\hat{p}(x_i\mid\omega_j)\hat{P}(\omega_j)}{\hat{p}(x_i\mid\omega_1)\hat{P}(\omega_1)+\hat{p}(x_i\mid\omega_2)\hat{P}(\omega_2)},\quad j=1,2。\tag{10}$$

利用式(10)即可估计出样本集中样本点 x_i 的后验概率 $P(\omega_j\mid x_i)$。

3.2 构造基于后验概率加权的隶属度函数

FSVM中的样本隶属度用于描述样本集中各样本点对于分类超平面的贡献大小,因此,在构造隶属度函数时,为体现出不同性质样本点之间的分类贡献差异,不仅要体现样本点在所在类中的重要程度,即体现出样本点属于所在类的可能性大小,还要尽量减小孤立点及噪声数据对生成分类超平面的作用。由贝叶斯决策规则可知,样本点对其所在类的后验概率不仅可以描述样本点属于所在类的可能性大小,同时还可以避免对孤立点和噪声数据的检测。

文献[8]利用样本的后验概率来表示属于所在类的可能性,将贝叶斯决策规则与SVM结合,建立了基于后验概率的支持向量机。但这种方法同样没有考虑类边界样本的特有属性。

对于给定的样本集,如果某个样本点周围的样本数越多,显然该样本点的样本密度就越大,则该样本点对分类的影响就越大。反之,如果样本点的样本密度越小,则该样本点对分类的影响就越小。

基于以上的分析,考虑对反映了样本所属类可能性的后验概率用密度比值 $\frac{\rho_i}{\rho}$ 进行加权,进而构造基于后验概率加权的隶属度函数。

定义5 样本集 T 中样本点 x_i 的基于后验概率加权的隶属度函数 $\mu(x_i)$ 定义为

$$\mu(x_i)=P(\omega_j\mid x_i)\cdot\left(\frac{\rho_i}{\rho}\right),j=1,2,i=1,2,\cdots,l。\tag{11}$$

其中:$\mu(x_i)$ 表示样本点 x_i 属于 ω_j 类的隶属度,$P(\omega_j\mid x_i)$ 表示样本点 x_i 对于所属类 ω_j 的后验概率,ρ 和 ρ_i 分别表示样本集 T 的平均密度和样本点 x_i 的样本密度。

在定义5中,后验概率 $P(\omega_j\mid x_i)$ 描述了样本点 x_i 属于类 ω_j 的可能性,ρ 和 ρ_i 是没加区分样本类别的整个样本集 T 的平均密度和样本密度,用 $\frac{\rho_i}{\rho}$ 对 $P(\omega_j\mid x_i)$ 进行加权以获得样本点的隶属度,是基于这样的思想:

①考虑到支持向量是分布于分类超平面附近,因而超平面附近的样本往往对分类影响较大,需要关注这些样本的位置分布。通过 $\frac{\rho_i}{\rho}$ 可以调节分类超平面附近样本对分类的作用。

②由贝叶斯公式和确定后验概率的经验方法可知,$P(\omega_j\mid x_i)$ 是样本点 x_i 在类 ω_j 中一个小邻域中局部的样本情况的描述,其值的大小,反映了样本点 x_i 属于类 ω_j 的可能性。而 $\frac{\rho_i}{\rho}$ 的值反映了样本点 x_i 在整个样本集 T 中的整体分布情况,尤其是对分类超平面附近的样本,当 $\frac{\rho_i}{\rho}$ 对 $P(\omega_j\mid x_i)$ 加权时,充分考虑了其对分类的作用。

③在贝叶斯决策中,生成决策规则的有效性和可靠性是依赖于对样本密度函数的准确估计。然而模糊支持向量分类机与贝叶斯决策毕竟不同,尽管也期望获得尽可能准确的概率估计,用前面经验估计的方法即使有一定的偏差,但在FSVM决策分类超平面设计中的Lagrange乘子 α_i 可调节样本点 x_i 对超平面的影响。同时,在估计类条件概率 $\hat{p}(x_i\mid$

ω_j)时,不同样本点的邻域可能会有重叠,使得样本集总的概率会超过 1,即 $\sum_i \hat{p}(x_i \mid \omega_j) > 1$,然而类条件概率估计的最终目的是进一步估计样本点的后验概率,类条件概率的估计仅仅是其中的一个中间步骤。

④由于 ρ 是样本集 T 的平均密度,使得 $\frac{\rho_i}{\rho}$ 的取值可能大于 1 也可能小于 1,当 $\frac{\rho_i}{\rho}<1$ 时,样本点 x_i 是孤立点或噪声点的可能性很大[12],此时用 $\frac{\rho_i}{\rho}$ 对 $P(\omega_j|x_i)$ 进行加权将减小 x_i 的隶属度,进而减弱其对分类的作用,避免了对异常数据的监测和对异常数据单独定义隶属度。但是对分类超平面附近的样本点,由于 ρ 和 ρ_i 是整个样本集上定义的,$P(\omega_j|x_i)$ 是同类样本的估计,因此 $\frac{\rho_i}{\rho}$ 加权又将增加这些样本的隶属度,进而增加其对分类的作用。

设有训练样本集

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\},$$

其中 $x_i\in\mathbf{R}^n$,$y_i\in\{-1,+1\}$,$i=1,2,\cdots,l$,考虑样本集中的孤立点或噪声点等不确定因素,用 FSVM 分类求解,此时将样本集进行改造如下:

$$(x_1,y_1,\mu_1),(x_2,y_2,\mu_2),\cdots,(x_l,y_l,\mu_l) \quad (12)$$

其中:y_i 表示分类状态标记,$y_i=+1$ 代表正类,用 ω_1 表示(记 ω_1 为+);$y_i=-1$ 代表负类,用 ω_2 表示(记 ω_2 为−)。隶属度 μ_i 由式(11)可知为

$$\mu_i=\mu(x_i)=\begin{cases}P(+|x_i)\cdot\left(\frac{\rho_i}{\rho}\right), y_i=+1,\\ P(-|x_i)\cdot\left(\frac{\rho_i}{\rho}\right), y_i=-1。\end{cases} \quad (13)$$

3.3　基于后验概率加权的隶属度求解算法

在模糊支持向量机建模过程中,当把样本空间中的样本通过映射 $\Phi(x)$ 映射到特征空间时,样本点之间的空间关系发生了变化,在利用式(13)计算样本点隶属度值时,也将通过特征空间进行计算。

由式(13),得到基于后验概率加权的隶属度求解算法如算法 1。

算法 1　基于后验概率加权的隶属度求解算法

输入:训练样本集

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\},$$

其中 $x_i\in\mathbf{R}^n$,$y_i\in\{-1,+1\}$,$i=1,2,\cdots,l$,给定 $0<\lambda<1$ 值。

输出:样本 x_i 的隶属度 μ_i,$i=1,2,\cdots,l$。

Step 1:由式(7)建立样本集 T 的距离矩阵 R,由式(3)~(4)计算样本集的平均距离 D 和最大距离 d。

Step 2:由式(5)计算样本集的平均密度 ρ。

Step 3:依据样本集 T 中 y_i 的值,由式(8)计算样本类先验概率 $\hat{P}(+)$和 $\hat{P}(-)$。

Step 4:对样本集 T 中每一个样本点 x_i,$i=1,2,\cdots,l$

{

1)扫描距离矩阵 $\boldsymbol{R}$,计算距 x_i 距离小于 $\sqrt{3}\lambda D/2$ 的样本数 l_i 和与 y_i 值同类的样本数 k_i^j。

2)由式(6)计算样本点 x_i 的密度 ρ_i。

3)$y_i=+1$ 时,由式(9)计算类条件概率 $\hat{p}(x_i|+)$,由式(10)计算样本点 x_i 的后验概率 $P(+|x_i)$;$y_i=-1$ 时,由式(9)计算类条件概率 $\hat{p}(x_i|-)$,由式(10)计算样本点 x_i 的后验概率 $P(-|x_i)$。

4)将 ρ、ρ_i、$P(+|x_i)$(或 $P(-|x_i)$)代入式(13),计算 μ_i。

}

算法 1 中,当样本集 T 中所有样本点的隶属度求出后算法结束。算法的计算时间消耗主要在求样本集中样本点之间的距离和对距离矩阵的扫描检索上,T 中样本总数为 l,故其时间复杂度是 $O(l^2)$。

4　仿真实验

为验证基于后验概率加权的模糊隶属度构造的有效性,通过建立基于改进隶属度的 FSVM,并应用到实例数据进行仿真实验。

仿真实验使用的环境:在 Windows XP 平台上用 Matlab7.1 的 SVM-KM 工具箱开发实现的,使用一台 1G 内存 Intel 酷睿双核 2.0GHz 处理器的计算机上完成。

4.1　仿真数据样本的获取

使用双螺旋线样本和 UCI 机器学习数据库[13]中的心脏扫描统计数据集(SPECT Heart Data Set)进行实验。

4.1.1　双螺旋线样本获取

双螺旋线样本集按下式产生[14]:

```
angle=(i*Π)/(16*density);
radius= maxRadius * ((104 * density − i)/(104 * density));
x=radius * cos(angle);
y=radius * sin(angle);    //第 1 条螺旋线
x=(−radius * cos(angle)−0.5);
y=(−radius * sin(angle)−0.5);    //第 2 条螺旋线。
```

利用上式产生如下样本集：

Data1：令 maxRadius＝3，density＝1，i 等于 1 到 100。2 条螺旋线共产生 200 个样本点，其中两条螺旋线分别随机产生 66 个共 132 个样本点作为训练样本集，余下 68 个样本点作为测试样本集。

4.1.2 SPECT Heart 数据集获取

UCI 机器学习数据库中的 SPECT Heart 数据集是断层扫描心脏的数据集，用于判断心脏是否正常。SPECT Heart 数据集包含有 267 个样本，其中用于训练样本 80 个，用于测试样本 187 个。实际上 SPECT Heart 数据集又分为两个子集：SPECT Heart 数据子集和 SPECTF Heart 数据子集，由此获得如下样本集：

Data2：SPECT Heart 数据子集，每个样本点有 23 个属性，其中属性 1 作为每个样本点的类属性，决定了样本点所属的类别，第 2 到第 23 个属性是取值为 0 或 1 的样本点特征值。值为 1 表示此样本点为正常，值为 0 时此样本点为不正常。

Data3：SPECTF Heart 数据子集，每个样本点有 45 个属性，其中第 2 到第 45 个属性是取值为 50～90 的样本点特征值。

4.2 基于后验概率加权隶属度的 FSVM 建立

模糊支持向量机的核心思想是引入隶属度函数对样本加权，扩展了传统 SVM 算法以解决不确定性分类问题[15-16]。这样可以减小孤立点和噪声点的影响，提高 SVM 的分类性能。

与传统支持向量机中的惩罚项不同，模糊支持向量机用训练样本的隶属度模糊化这个惩罚项。训练样本的隶属度 μ_i 由样本特征的后验概率加权即式(13)来确定，从而建立起模糊支持向量机最优超平面的优化问题[9-10]。通过建立模糊支持向量机优化问题，进而验证本文建立的基于后验概率加权模糊隶属度构造的有效性。

4.3 仿真实验结果

对双螺旋线样本集 Data1 和 SPECT Heart 数据集 Data2、Data3 分别用标准的 C-SVM、基于线性隶属度函数的 FSVM[3] 和本文基于后验概率加权隶属度的 FSVM 进行训练学习，得到相应的分类决策函数。

实验中惩罚系数 C＝50，核函数采用 Gauss 径向基核函数，利用相应决策分类函数对测试样本集进行测试，获得不同样本集分类测试精度的比较如表 1。

表 1 3 种不同支持向量分类机测试精度比较

SVM 算法	数据集	训练集正类样本数	训练集负类样本数	测试集正类样本数	测试集负类样本数	正类测试精度/%	负类测试精度/%	测试精度/%
标准 C-SVM	Data1	66	66	34	34	100	50	75
线性函数 FSVM[3]	Data1	66	66	34	34	100	50	75
后验概率加权隶属度 FSVM	Data1	66	66	34	34	55.88	100	77.94
标准 C-SVM	Data2	40	40	172	15	57.56	33.33	55.61
线性函数 FSVM[3]	Data2	40	40	172	15	48.26	66.67	49.73
后验概率加权隶属度 FSVM	Data2	40	40	172	15	79.07	60	77.54
标准 C-SVM	Data3	40	40	172	15	49.42	53.33	49.73
线性函数 FSVM[3]	Data3	40	40	172	15	63.37	46.67	62.03
后验概率加权隶属度 FSVM	Data3	40	40	172	15	69.19	53.33	67.91

从表 1 可以看出，笔者提出的基于后验概率加权隶属度函数构造方法，通过建立 FSVM 分类学习训练，在同一数据样本集上，其分类精度高于传统的 C-SVM 和基于线性函数的 FSVM，在不同数据集上表现出了较好的分类能力。

5 结 论

受贝叶斯决策的启发，依据不同性质样本点在整个样本集中的相对位置分布，引入样本点的后验概率和样本点密度，通过数据样本集的后验概率经验估计，利用后验概率对样本密度进行加权，提出了后验概率加权的隶属度函数构造方法。该方法不单

独对孤立点和噪音数据进行检测，以统一的后验概率加权设计隶属度函数，一方面可以减少孤立点和噪音数据对分类超平面的影响，另一方面，不影响支持向量对分类超平面的决定作用。

参考文献：

[1] 顾亚祥，顶世飞. 支持向量机研究进展[J]. 计算机科学，2011，38(2)：14-17.
GU Yaxiang，DING Shifei. Advances of support vector machines [J]. Computer Science，2011，38(2)：14-17.

[2] Liu C F，Wang S D. Fuzzy support vector machines [J]. IEEE Transactions on Neural Networks，2002，13(2)：464-471.

[3] 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.

[4] 边肇祺，张学工. 模式识别[M]. 第二版. 北京：清华大学出版社，2000.

[5] 唐浩，廖与禾，孙峰，等. 具有模糊隶属度的模糊支持向量机算法[J]. 西安交通大学学报，2009，43(7)：40-43.
TANG Hao，LIAO Yuhe，SUN Feng，et al. FSVM with a new fuzzy membership function [J]. Journal of Xi'an Jiaotong University，2009，43(7)：40-43.

[6] Li M M，Xiang F H，Liu X W. A novel membership function for fuzzy support vector machines [J]. Computer Engineering & Science，2009，31(9)：92-94.

[7] Li L，Zhou M M，Lu Y L. Fuzzy support vector machine based on density with dual membership [J]. Computer Technology and Development，2009，19(12)：44-46.

[8] 吴高巍，陶卿，王珏. 基于后验概率的支持向量机[J]. 计算机研究与发展，2005，42(2)：196-202.
WU Gaowei，TAO Qing，WANG Jue. Support vector machines based on posterior probability [J]. Journal of Computer Research and Development，2005，42(2)：196-202.

[9] 魏延，石磊，陈琳琳. 基于后验概率加权的模糊支持向量机[J]. 重庆工学院学报：自然科学版，2009，23(8)：80-84.
WEI Yan，SHI Lei，CHEN Linlin. Fuzzy support vector machine based on posterior probability weight [J]. Journal of Chongqing Institute of Technology：Natural Science，2009，23(8)：80-84.

[10] 石磊. 基于后验概率加权的模糊支持向量分类机研究及应用[D]. 重庆：重庆师范大学，2009.

[11] Zhang Y，Chi Z X. A fuzzy support vector classifier based on Bayesian optimization [J]. Fuzzy Optimization and Decision Making，2008，7(1)：75 - 86.

[12] 施化吉，周书勇，李星毅，等. 基于平均密度的孤立点检测研究[J]. 电子科技大学学报，2007，36(6)：1286-1288.
SHI Huaji，ZHOU Shuyong，LI Xinyi，et al. Average density-based outlier detection [J]. Journal of University of Electronic Science and Technology of China，2007，36(6)：1286-1288.

[13] Frank A，Asuncion A. UCI machine learning repository[EB/OL]. [2012-02-01]. http://archive.ics.uci.edu/ml/datasets.html.

[14] Singh S. Neural learning of spiral structures[C]// Proceedings of the International Conference on Advances in Pattern Recognition，November 23-25，1998，Plymouth，UK. New York：Springer，1998：226-231.

[15] Tang H，Qu L S. Fuzzy support vector machine with a new fuzzy membership function for pattern classification [C] // Proceedings of the Seventh International Conference on Machine Learning and Cybernetics，July 12-15，2008，Kunming，China. [S. l.]：Institute of Electrical and Electronics Engineers Computer Society，2008，2：768-773.

[16] Nemmour H，Chibani Y. Fuzzy integral to speed up support vector machines training for pattern classification [J]. International Journal of Knowledge-Based Intelligent Engineering Systems，2010，14(3)：127-138.

（编辑　张小强）