一种新的模糊支持向量机

哈明虎,彭桂兵,赵秋焕,马丽娟
HA Ming-hu,PENG Gui-bing,ZHAO Qiu-huan,MA Li-juan

河北大学 数学与计算机学院,河北 保定 071002
College of Mathematics and Computer Science,Hebei University,Baoding,Hebei 071002,China
E-mail:mhha@mail.hbu.edu.cn

HA Ming-hu,PENG Gui-bing,ZHAO Qiu-huan,et al.New fuzzy support vector machine.Computer Engineering and Applications,2009,45(25):151-153.

Abstract: Fuzzy Support Vector Machine(FSVM),which are the design methods of membership functions are based on class-center,and can effectively overcome the problem that the Support Vector Machine(SVM) is sensitive to the noises and outliers; however,it assigns smaller memberships to the support vectors,which may decrease the effects of these support vectors to the construction of the classification hyperplane.To tackle the above problem,a novel method to determine membership function is proposed.At the same time,the training time of FSVM is generally long which is aroused by the high computational complexity of constructing its kernel function.To reduce the training time of FSVM,the training samples are clustered by an effective sectional set fuzzy C-means clustering(S2FCM) firstly.Then,the cluster centers are taken as training samples.According to the novel method to determine membership function and the S2FCM,a new FSVM is constructed.Experimental results show that the new FSVM can effectively enhance the training speed and classification accuracy rate.

Key words: fuzzy support vector machine;membership function;sectional set fuzzy C-means clustering

摘 要:基于类中心设计隶属度函数的模糊支持向量机能有效地解决支持向量机对噪声或孤立点敏感度高的问题,但是,由于它对支持向量赋予较小的隶属度,从而降低了其分类作用。基于此,提出一种新的隶属度函数设计方法;同时,针对模糊支持向量机普遍存在因核函数计算量大,而导致训练时间长的问题,通过使用一种高效的截集模糊C-均值聚类方法对训练样本进行聚类,然后以聚类中心作为样本进行训练,以减少训练样本来提高训练速度。根据上述新的隶属度函数设计方法和截集模糊C-均值聚类方法,构建了一种基于截集模糊C-均值聚类并改进了隶属度函数的模糊支持向量机,数值试验表明这种新的模糊支持向量机有效地提高了训练速度和分类精度。

关键词:模糊支持向量机;隶属度函数;截集模糊C-均值聚类
DOI:10.3778/j.issn.1002-8331.2009.25.046 **文章编号**:1002-8331(2009)25-0151-03 **文献标识码**:A **中图分类号**:TP391

支持向量机(Support Vector Machine,SVM)是一种基于统计学习理论的通用机器学习方法[1-2],也是数据挖掘中的一项新技术[3],最初于20世纪90年代由Vapnik等人提出。它使用结构风险最小化原则代替经验风险最小化原则,能较好地处理小样本情况下的学习问题[4];同时采用了核函数思想,能把非线性问题转化为线性问题来解决。近来,由于SVM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。但SVM目前还存在一定的局限性,例如对训练样本内的噪声或孤立点反应敏感[5],对不是完全属于两类中一类的样本分类正确率不高。针对这些不足,Lin等学者[6]将隶属度的概念引入到支持向量机,构建了模糊支持向量机(Fuzzy Support Vector Machine,FSVM)。模糊支持向量机,是在支持向量机的基础上给每个样本分别赋一个隶属度值,对不同的样本采用不同的惩罚权重系数,在构造目标函数时,使不同的样本有不同的贡献,对噪声或孤立点赋予很小的权值,从而达到消除噪声或孤立点的目的。

在模糊支持向量机中,隶属度函数的设计是关键。Lin等[6]给出了一种基于类中心的隶属度函数设计方法,该方法简单易

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60773062);教育部科学技术研究重点项目计划(the Key Scientific and Technical Research Project of the Ministry of Education of China.No.206012); 河北省自然科学基金(the Natural Science Foundation of Hebei Province of China under Grant No.2008000633); 河北省省教育厅科研计划重点项目 (the Key Scientific Research Project of Department of Hebei Education of China under Grant No.2005001D)。

作者简介:哈明虎(1963-),男,教授,主要研究方向:不确定信息处理,不确定统计预测与决策等;彭桂兵(1982-),男,硕士生,主要研究方向:不确定支持向量机;赵秋焕(1983-),女,硕士生,主要研究方向:不确定支持向量机;马丽娟(1982-),女,硕士生,主要研究方向:不确定支持向量机。

收稿日期:2009-05-11 **修回日期**:2009-06-18

行,但在减小噪声或孤立点作用的同时对支持向量赋予较小的隶属度，降低了支持向量对分类的贡献；同时对非均衡的样本，有时并不能很好地将噪声或孤立点从有效样本集中分离出来[7]。为此,提出了一种新的隶属度函数设计方法,不但提高了支持向量的隶属度,也考虑到了非均衡样本,对分离噪声或孤立点更有效。

模糊支持向量机能够提高支持向量机的分类精度,但仍然因其所需内存需求大,核函数计算量大,训练时间长等不足而使应用不便。因此如何优化模糊支持向量机的训练速度就成为一个重要问题,本文应用一种新的高效软聚类方法—截集模糊C-均值聚类(Sectional Set Fuzzy C-Means,S2FCM)[8]对样本个数进行优化约简来提高运行速度。

截集模糊C-均值聚类是模糊C-均值聚类（Fuzzy C-Means,FCM)的一个改进[9]。FCM是一种高效的"软聚类"算法,它允许一个样本以不同的隶属度属于所有不同的类。但在实际问题中，当一个样本属于某类的隶属度远大于属于其他类的隶属度时,应用最大隶属度原则,可以将其划为该类,不必再考虑其属于其他类的情况，只有对于属于各个类的隶属度都很接近的样本,才将其以隶属度与各类联系。为此,引入了截集模糊C-均值聚类算法,此算法可以加快聚类速度并使聚类更加有效合理。

基于上述新的隶属度函数设计方法和截集模糊C-均值聚类算法,构建了一种新的模糊支持向量机。

1 一种新的隶属度函数设计方法

为了减小噪声或孤立点对分类超平面的影响,Lin等学者在文献[6]中提出了一种基于类中心的隶属度函数设计方法,使样本对分类所起的作用随着样本远离类别的几何中心而逐渐减小,这样可以弱化噪声或孤立点的影响。但是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来确定的,而支持向量往往都离类别中心较远,按照文献[6]中的方法设计隶属度函数,在减小噪声或孤立点对分类超平面作用的同时,也大大弱化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。如图1所示,从圆心到圆周,训练样本的隶属度依次减小,这样,支持向量(粗体表示)的隶属度很小,从而容易导致分类超平面偏离最优分类超平面。

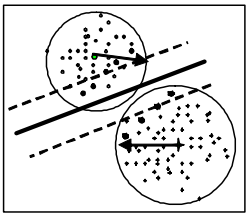

图1 基于类中心的隶属度设计方法

给出了一种新的隶属度函数设计方法,使样本对分类所起的作用随着样本远离类别的几何中心而逐渐增大,即将样本到类别几何中心的距离与该类中离类别几何中心最远的样本到类别几何中心的距离的比值定义为隶属度。但是,当样本与类别几何中心的距离大于阈值时，就给样本赋一个很小的隶属度,阈值是根据两类样本几何中心之间的距离和样本的稠密情况决定的。这样通过调整阈值,就可以使支持向量的隶属度较大,而噪声或孤立点的隶属度很小。如图2所示,从圆心到虚线圆周,训练样本的隶属度依次增大,而虚线圆周与实线圆周之间的训练样本则赋予很小的隶属度,这样,在给噪声或孤立点很小隶属度的同时,保证了支持向量(在虚直线上)有较大的隶属度,从而使分类精度较高。

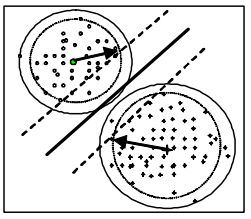

图2 新的隶属度设计方法

使用正类样本的均值作为正类的中心,记为 x^+,即 $x^+=1/l^+\sum_{i=1}^{l^+}x_i$;负类样本的均值作为负类的中心,记为 x^-,即 $x^-=1/l^-\sum_{i=1}^{l^-}x_i$。正类的半径 $r^+=\max_{i=1,2,\cdots,l^+}\|x_i-x^+\|$,负类的半径 $r^-=\max_{i=1,2,\cdots,l^-}\|x_i-x^-\|$。每个正类样本到正类中心的距离为 $d_i^+=\|x_i-x^+\|$,并计算平均距离 $m^+=1/l^+\sum_{i=1}^{l^+}d_i^+$;每个负类样本到负类中心的距离为 $d_i^-=\|x_i-x^-\|$,并计算平均距离 $m^-=1/l^-\sum_{i=1}^{l^-}d_i^-$。两类中心的距离为 $t=|x^+-x^-|$,θ 为一个事先给定的很小的正数,作为噪声和孤立点的隶属度。$\lambda>0$ 为一个控制因子,使 $t\cdot\lambda\cdot m^+/(m^++m^-)<r^+$ 和 $t\cdot\lambda\cdot m^-/(m^++m^-)<r^-$ 成立。则隶属度函数定义为:

$$S_i^+=\begin{cases}\dfrac{\delta+d_i^+}{r^+}, & d_i^+\leqslant t\cdot\lambda\cdot\dfrac{m^+}{m^++m^-}\\ \theta, & d_i^+>t\cdot\lambda\cdot\dfrac{m^+}{m^++m^-}\end{cases}$$

$$S_i^-=\begin{cases}\dfrac{\delta+d_i^-}{r^-}, & d_i^-\leqslant t\cdot\lambda\cdot\dfrac{m^-}{m^++m^-}\\ \theta, & d_i^->t\cdot\lambda\cdot\dfrac{m^-}{m^++m^-}\end{cases}\tag{1}$$

式中 δ 是足够小的正数,为了保证 $s_i>0$。

2 截集模糊C-均值聚类算法

设数据集合 $X\subset R^{s\times n}$ 的模糊C-划分 $U=[u_{ik}]_{c\times n}$,$(1\leqslant i\leqslant c,1\leqslant k\leqslant n)$及 $\lambda\in[0,1]$,令

$$U_{pk}=\max\{U_{ik},1\leqslant i\leqslant c\}\tag{2}$$

$$w_{ik}=\begin{cases}1,U_{pk}\geqslant\lambda \text{ and } i=p\\ 0,U_{pk}\geqslant\lambda \text{ and } i\neq p\\ u_{ik},U_{pk}<\lambda,\forall\,1\leqslant i\leqslant c\end{cases}\tag{3}$$

则 $W=[w_{ik}]_{c\times n}$ 为数据集合 $X\subset R^{s\times n}$ 的 λ 截集模糊C-划分[9]。

依据Bezdek的ISODATA算法来构建截集模糊C-均值聚类算法(S2FCM),具体步骤如下[8]:

初始化指数因子 m,停止误差 ε,分类类数 c,截集因子 $\lambda=0.5+1/(2c)$。

步骤1 在数据集合 $X=\{x_1,x_2,\cdots,x_n\}$,随机选择 c 个数据作为初始聚类中心 $V=\{v_1,v_2,\cdots,v_c\}$[10];

步骤2 计算 $x_k(1\leqslant k\leqslant n)$到聚类中心 $v_i(1\leqslant i\leqslant c)$的内积范数:$d_{ik}=\|x_k-v_i\|_A^2$;

步骤3 令 $d_{c+1k}=\min\{d_{ik},1\leqslant i\leqslant c\}(1\leqslant k\leqslant n)$,返回 d_{c+1k} 值对应的 i 的值,记作 s,并计算 $p_k=\left(\dfrac{1}{d_{c+1k}}\right)^{\frac{1}{m-1}}\Big/\sum_{i=1}^{c}\left(\dfrac{1}{d_{ik}}\right)^{\frac{1}{m-1}}$。对每

一个 $x_k(1\leqslant k\leqslant n)$进行如下处理:

如果 $d_{c+1k}=0$, 或者 $d_{c+1k}\neq 0$ 但 $p_k\geqslant\lambda$, 那么 $w_{ik}=\begin{cases}1 & i=s\\ 0 & i\neq s,\forall 1\leqslant i\leqslant c\end{cases}$;

如果 $d_{c+1k}\neq 0$,且 $p_k<\lambda$,那么 $w_{ik}=\left(\frac{1}{d_{ik}}\right)^{\frac{1}{m-1}}/\sum_{i=1}^{c}\left(\frac{1}{d_{ik}}\right)^{\frac{1}{m-1}}$。

通过上面的计算得到 $W=[w_{ik}]_{c\times n}$;

步骤 4 用非零的 w_{ik} 计算 $v_i=\sum_{k=1}^{n}(w_{ik})^m x_k/\sum_{k=1}^{n}(w_{ik})^m$,得到新的聚类中心 V^{l+1};

步骤 5 如果 $\|V^{l+1}-V^l\|^2\leqslant\varepsilon$,迭代终止,并且输出聚类中心 V^{l+1},否则令 $l=l+1$,返回步骤 2。

在 S2FCM 算法中,当 $\lambda>0.5$ 时,在式(3)的条件下某些列非零 w_{ik} 只有一个,计算隶属度时,只要出现满足式(3)的非零 w_{ik},则可以马上得到该样本隶属度,且计算过程中均为稀疏矩阵,所以具有很快的收敛速度,与模糊支持向量机核函数计算量相比具有明显优势。

3　新的模糊支持向量机

主要讨论两类分类的情况。假设训练样本表示为:

$(x_1',y_1'),(x_2',y_2'),\cdots,(x_n',y_n')$

每个样本的特征表示为 $x_i'\in\boldsymbol{R}^n$, 类标识为 $y_i'=\{-1,1\},i=1,2,\cdots,n$。

使用截集模糊 C-均值聚类算法对样本进行聚类, 得到新的训练样本

$(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)$

用隶属度函数设计方法计算出各个样本的隶属度 $s_i(0<s_i\leqslant 1,i=1,2,\cdots,l)$,它表示第 i 个样本属于正常的程度,从而得到模糊支持向量机的训练样本

$(x_1,y_1,s_1),(x_2,y_2,s_2),\cdots,(x_l,y_l,s_l)$

其中 $x_i\in\boldsymbol{R}^n$,类标识为 $y_i=\{-1,1\}$,隶属度为 $s_i,i=1,2,\cdots,l$。假设 $z=\varphi(x)$ 为训练样本从输入空间 $\boldsymbol{R}^n$ 到高维特征空间 $\boldsymbol{Z}$ 的映射关系。

由于隶属度 s_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 $s_i\xi_i$ 为带权的误差项,由此得到最优分类面为下面目标函数的最优解

$$\begin{aligned}&\min\ \Phi(w,\xi)=\frac{1}{2}w\cdot w+C\left(\sum_{i=1}^{l}s_i\xi_i\right)\\&\text{s.t. } y_i[(w\cdot z_i)+b]-1+\xi_i\geqslant 0,i=1,2,\cdots,l\\&\quad\xi_i\geqslant 0,i=1,2,\cdots,l\end{aligned}\tag{4}$$

其中,惩罚因子 C 为常数,s_i 越小,则相应的样本 x_i 在对式(4)优化问题所起的作用就越小。为求目标函数的最优解,构造拉格朗日函数

$$L(w,b,\xi,\alpha,\beta)=\frac{1}{2}w\cdot w+C\sum_{i=1}^{l}s_i\xi_i-\sum_{i=1}^{l}\alpha_i(y_i(w\cdot z_i+b)-1+\xi_i)-\sum_{i=1}^{l}\beta_i\xi_i\tag{5}$$

其中,$\alpha_i,\beta_i\geqslant 0$ 为 Lagrange 乘子。

在鞍点处变量满足:

$$\frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial w}=w-\sum_{i=1}^{l}\alpha_i y_i z_i=0\tag{6}$$

$$\frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial b}=-\sum_{i=1}^{l}\alpha_i y_i=0\tag{7}$$

$$\frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial \xi_i}=s_iC-\alpha_i-\beta_i=0\tag{8}$$

将这些条件代入到式(5),得到原问题式(4)的对偶规划为:

$$\begin{aligned}&\max\ W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\\&\text{s.t. }\sum_{i=1}^{l}\alpha_i y_i=0,0\leqslant\alpha_i\leqslant s_iC\quad i=1,2,\cdots,l\end{aligned}\tag{9}$$

上式中 $K(x_i,x_j)=z_i\cdot z_j=\varphi(x_i)\cdot\varphi(x_j)$为核函数。Kuhn-tucker 条件为:

$$\begin{aligned}&\alpha_i(y_i(w\cdot z_i+b)-1+\xi_i)=0,i=1,2,\cdots,l\\&(s_iC-\alpha_i)\xi_i=0,i=1,2,\cdots,l\end{aligned}\tag{10}$$

相应的决策函数为:

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b\right)\tag{11}$$

$\alpha_i=0$ 对应的样本 $\boldsymbol{x}_i$ 是被完全正确分类的,$\alpha_i>0$ 对应的样本 $\boldsymbol{x}_i$ 被称为支持向量。这里有两种类型的支持向量,一种满足 $0<\alpha_i<s_iC$ 的支持向量 $\boldsymbol{x}_i$ 位于分类超平面附近;另一种满足 $\alpha_i=s_iC$ 的支持向量 $\boldsymbol{x}_i$ 为被误分类的样本。模糊支持向量机方法和支持向量机方法的差别在于,模糊支持向量机中含有隶属度 s_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。

从上面的公式可以看出,s_iC 表示样本 $\boldsymbol{x}_i$ 在训练支持向量机时的重要程度。s_iC 越大,表示样本 $\boldsymbol{x}_i$ 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,s_iC 越大,表示样本 $\boldsymbol{x}_i$ 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。 对于孤立点或噪声样本, 如果能够使其对应的 s_i 很小,从而 s_iC 很小,则此样本对支持向量机的训练所起的作用就大为减小了,其结果便是大大降低它们对分类面的影响。

4　数值实验

对真实数据集 Splice 和 ijcnn1 进行实验, 分别采用 SVM 算法,文献[6]中提出的 FSVM 算法,以及该文中提出的新的 FSVM(NFSVM)算法。在相同的参数条件下,分别运行 10 次,取平均值,比较结果见表 1。

表 1　三种算法的运行时间和精确度比较

数据集(特征个数)(训练/测试)	算法	训练数据个数	运行时间/s	分类精度/(%)
Splice(60)(1 000/2 175)	SVM	1 000	2.92	79.862 1
	FSVM	1 000	4.06	83.524 7
	NFSVM	225	2.16	84.267 3
ijcnn1(22)(49 990/91 701)	SVM	49 990	118.97	86.531 0
	FSVM	49 990	152.56	86.972 6
	NFSVM	12 000	76.17	87.063 7

5　结论

在基于类中心的隶属度函数设计方法的基础上,提出了一种新的隶属度函数设计方法,使用该方法可以给噪声或孤立点赋予很小的隶属度,以降低其对分类面的影响,同时又能确保支持向量有较大的隶属度,从而提高分类精度;使用截集模糊 C-均值聚类算法对样本进行有效的聚类, 以聚类中心作为训练样本,以减少样本数量来提高训练速度。将新的隶属度函数设计方法和截集模糊 C-均值聚类算法应用到模糊支持向量机中,形成了新的模糊支持向量机。通过实验对比分析,提出的新模糊支持向量机,在训练速度和精度上都有了一定的提升。

(下转 194 页)